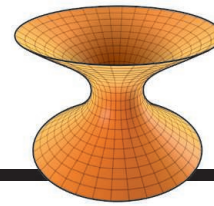# Science

AAAS

CLOSING WITH
SATURN

Cassini explores within the
planet's rings *p. 44*

CONTENTS

5 OCTOBER 2018 • VOLUME 362 • ISSUE 6410

32 & 65

Oscillations in
an insulator

25 & 72

CREDITS: (GRAPHIC) C. BICKEL/*SCIENCE*; (PHOTO) WESTEND61/GETTY IMAGES

NEWS

Published by AAAS

CONTENTS

28 & 69

Probing plasmonic catalysis

CLOSING WITH SATURN

ON THE COVER

Cassini's final wide-angle view of Saturn and its rings, taken on 13 September 2017, 2 days before the spacecraft burned up in the planet's atmosphere. The Sun is at an oblique angle on the far side of Saturn, which casts a dark shadow across the rings. This color mosaic was produced from multiple images taken with Cassini's Imaging Science Subsystem. See page 44. *Image: NASA/JPL-Caltech/Space Science Institute*

Renewable energy for Puerto Rico

Puerto Rico is not prepared for another hurricane. A year ago, Hurricane María obliterated the island's electric grid, leading to the longest power outage in U.S. history. This disrupted medical care for thousands and contributed to an estimated 2975 deaths. The hurricane caused over $90 billion in damage for an island already in economic crisis. Although authorities claim that power was restored completely, some residents still lack electricity. Despite recovery efforts, the continued vulnerability of the energy infrastructure threatens Puerto Rico's future. But disruptions create possibilities for change. Hurricane María brought an opportunity to move away from a fossil fuel–dominant system and establish instead a decentralized system that generates energy with clean and renewable sources. This is the path that will bring resilience to Puerto Rico.

Puerto Rico is representative of the Caribbean islands that rely heavily on fossil fuels for electric power; 98% of its electricity comes from imported fossil fuels (oil, natural gas, and coal), whereas only 2% comes from renewable sources (solar, wind, or hydroelectric). The distribution of 6023 MW is challenging, requiring thousands of miles of transmission and distribution lines over the island's steep topography. This makes the island's centralized electrical grid vulnerable to hurricanes that are predicted to increase in severity because of climate change.

In Puerto Rico and the rest of the Caribbean, where sun, wind, water, and biomass are abundant sources of renewable energy, there is no need to rely on fossil fuel technology. Unfortunately, the government of Puerto Rico and the U.S. Federal Emergency Management Agency have been making decisions about the local power authority that are restoring the energy system to what it was before Hurricane María hit, perpetuating fossil fuel reliance.

Despite these decisions, a transformation has begun in communities across Puerto Rico. For example, in the mountain municipality of Adjuntas, local initiatives headed by Casa Pueblo, a self-reliant nonprofit community organization, has increased the installation of solar energy systems. Fortunately, the solar power–based infrastructure of Casa Pueblo was not affected by the hurricane, allowing Adjuntas to serve as the organization's center of operations for immediate local and regional response after the hurricane. Adjuntas became an oasis of power, where people got immediate assistance. Analog solar-based energy systems were designed and installed by Casa Pueblo to supply the needs of numerous entities in the community: medical equipment, such as peritoneal dialysis for homes with patients; a radio transmitter for a community radio station; and equipment for hardware stores, mini-markets, restaurants, and other businesses. Around the island, other examples of off-the-grid local energy production reflect community resilience grounded in projects that foster renewable energy. They include a solar microgrid in Orocovis, multiple community aqueducts, and sustainable farms. These new energy systems are changing the energy landscape of the municipality. But the majority of rural communities is still in need of sustained help.

At this juncture, when the opportunity to build a sustainable and resilient electrical system presents itself, moving away from dependency on imported fossil fuels should be the guiding vision. Puerto Rico must embrace the renewable endogenous sources that abound on the island and build robust microgrids powered by solar and wind, install hybrid systems (such as biomass biodigesters), and create intelligent networks that can increase the resilience of the island. The Puerto Rican government and U.S. Congress should use Hurricane María as a turning point for pushing Puerto Rico toward using 100% renewable energy rather than a platform to plant generators across the island. The Fiscal Plans approved and certified by the Financial Oversight and Management Board for Puerto Rico, created by Congress in 2016, should be amended to pursue this vision of sustainable development based on renewable energy.

"...moving away from dependency on imported fossil fuels should be the guiding vision."

–**Arturo Massol-Deyá, Jennie C. Stephens, Jorge L. Colón**

Arturo Massol-Deyá *is a professor in the Department of Biology at the University of Puerto Rico, Mayagüez, Puerto Rico, and a member of Ciencia Puerto Rico. arturo.massol@upr.edu*

Jennie C. Stephens *is a professor of sustainability science and policy at Northeastern University and director for Strategic Research Collaborations at the Global Resilience Institute at Northeastern University, Boston, MA, USA. j.stephens@northeastern.edu*

Jorge L. Colón *is a professor in the Department of Chemistry at the University of Puerto Rico, San Juan, Puerto Rico, and a member of Ciencia Puerto Rico. jorge.colon10@upr.edu*

IN BRIEF | *Edited by* **Jeffrey Brainard**

SEISMOLOGY

Tsunami detectors fail to avert Indonesian deaths

A devastating 7.5-magnitude earthquake, followed by a tsunami wave reportedly up to 6 meters in height, swept into Palu Bay on Indonesia's Sulawesi island last week, killing at least 1234 people as of press time—a toll likely to rise as the disaster response continues. With a dearth of heavy equipment, many rescues in Palu, a city of 330,000, had to be attempted by hand, while a combination of soil liquefaction and landslides blocked roads to remote regions and collapsed homes. Researchers are now investigating why Palu reportedly did not receive warning of the tsunami, despite a new national alert system funded in part by Germany, the United States, and others. Contrary to some media reports, the tsunami warning system has not used seafloor sensors and buoys since 2010; instead it relies on seismometers and tide gauges. Five minutes after the earthquake, the warning system, which has been operated by Indonesia since 2011, predicted a wave of up to 3 meters for Palu, which should have triggered a mandatory evacuation along its shoreline, according to the GFZ German Research Centre for Geosciences in Potsdam, which helped design the system. Why this warning did not reach Palu's residents is now under investigation.

NSF limits biology proposals

RESEARCH FUNDING | Scientists are worried about a policy change announced in August by the National Science Foundation's (NSF's) biology directorate that allows them to send the directorate's three core tracks only one proposal a year in which they are listed as a principal investigator (PI) or co-PI. Although biologists can still seek funding from several other NSF programs unaffected by the rule, many are concerned that it will discourage collaborations and disproportionately affect early career scientists who depend on being listed as PIs for career advancement. Last week, 70 researchers signed a letter asking the agency to reconsider the new policy, which they complain was adopted without community input. NSF officials say the new limit was intended, in part, to reduce the number of rejected proposals resubmitted without major changes.

Climate mission moves ahead

EARTH SCIENCE | A space station sensor to monitor climate change is moving ahead, despite attempts by President Donald

Investigators will examine why a tsunami warning did not reach residents of Palu, Indonesia.

Trump's administration to kill it. Last week, NASA announced it had awarded a $57 million contract to the University of Colorado in Boulder to build the primary instrument of the Climate Absolute Radiance and Refractivity Observatory (CLARREO) Pathfinder. It was one of several earth science missions targeted by the administration for cancellation in budget proposals, only to be restored by Congress. To be installed on the International Space Station (ISS) early next decade, CLARREO Pathfinder will reduce uncertainties in measurements of reflected sunlight, helping calibrate other satellites, including those that gauge the planet's energy imbalance and track rainfall. Meanwhile, the agency announced last month that another ISS mission, the Orbiting Carbon Observatory-3, which will track the flow of greenhouse gases, has been completed and will launch in February 2019.

Argentina's science faces crisis

RESEARCH FUNDING | Argentina's economic crisis is wreaking havoc on science. As part of an austerity push, the government has proposed cutting the research budget from 3.7 billion pesos (roughly $96 million) this year to 3.4 billion pesos (about $88 million) in 2019; with inflation factored in, that's effectively a 35% cut, by one estimate. The National Scientific and Technical Research Council, which pays most researchers' salaries, will get a 27% increase, but even that isn't enough to keep up with the expected inflation rate for this year. Meanwhile, the government is far behind in its financial commitments to labs for 2018. Scientific organizations have also decried the government's decision to eliminate the Ministry of Science, Technology, and Productive Innovation, created in 2007, and make it part of the education ministry.

NIH awards go to more women

BIOMEDICINE | Women did unusually well this year in the latest round of the National Institutes of Health's prestigious high-risk, high-reward research awards. Half of this year's 10 Pioneer Awards went to women, compared with only one of 12 in 2017. And five of 11 Early Independence Award winners were female. For two other categories, the share that went to women was somewhat lower: They received 19 of 58 New Innovator Awards and were four of 18 Transformative Award winners. Expected to total $282 million over 5 years, the program selects investigators based on their track record and ideas, rather than a specific research project. The agency has been under pressure to ensure that the review process does not disadvantage female applicants.

Discord splits medical group

BIOMEDICINE | A rift within Cochrane, the world's premier organization for evidence-based medicine, is escalating. On 28 September, the director of the Nordic Cochrane Center in Copenhagen, Peter Gøtzsche, said he is withdrawing his center from the international collaboration, headquartered in London. Cochrane's governing board voted to expel Gøtzsche on 13 September for "a consistent pattern of disruptive and inappropriate behaviours," which led four other board members to resign in protest. Gøtzsche says Cochrane has since taken control of the website of the Copenhagen center, removed his statement about the case, and replaced it with one from the governing board. "The Great Leader, or Big Brother, in London is in control," says Gøtzsche, who has since posted a raft of documents—including a confidential report written by an external lawyer at Cochrane's request—on his own website,

80,000

Estimated U.S. deaths from influenza during the 2017–18 flu season, the highest in 4 decades (Centers for Disease Control and Prevention [CDC]).

35%

Portion of the world's freshwater and coastal wetlands lost to development and other causes from 1970 to 2015, a decline three times the rate of forest loss (Ramsar Convention on Wetlands).

918

Cases of congenital syphilis in the United States in 2017. The rate of such births has nearly tripled since 2012, in parallel with a growing number of cases in reproductive-aged women (CDC).

www.deadlymedicines.dk. Gøtzsche says he is backed by his center's staff and that he is convinced he can keep using the name "Cochrane" even after the split. A Cochrane spokesperson was unavailable to comment.

Physicist faulted for gender talk

GENDER | CERN, the European particle physics laboratory near Geneva, Switzerland, has suspended a guest researcher who gave a talk during a workshop there on gender issues in which he claimed that there is no discrimination against women in physics and complained that a woman had gotten a job he wanted even though his papers had more citations. Alessandro Strumia of the University of Pisa in Italy will be barred from any activities at the facility pending an investigation into the talk, which was "contrary to the CERN code of conduct," the organization announced on 1 October. The university said on the same day that it would also open an ethics investigation into the matter. CERN said workshop organizers were not aware of the content of the talk beforehand. Strumia declined to comment on the investigations.

U.S. quantum strategy unveiled

QUANTUM COMPUTING | U.S. scientific, business, and government leaders met last week at the White House to brainstorm a national strategy for quantum information science (QIS), such as developing quantum computers that could perform calculations that overwhelm an ordinary computer. The United States is playing catch-up: Last year, the European Commission launched a €1 billion, 10-year QIS initiative, and China is building a $10 billion QIS laboratory in Hefei. At the meeting, the U.S. National Science Foundation announced $31 million in grants for quantum research, the Department of Energy announced $218 million in funding for QIS, and the National Institute for Standards and Technology said it was forming a QIS consortium. Meanwhile, the U.S. House of Representatives last month approved a National Quantum Initiative that would provide $1.3 billion over 5 years for research. The Senate has yet to act on similar legislation.

EPA science advice downgraded

SCIENCE AGENCIES | A plan meant to streamline the U.S. Environmental Protection Agency's (EPA's) bureaucracy is again raising concerns that President Donald Trump's administration has little interest in scientific expertise. EPA officials last week said they will combine the agency's offices for science advice and science policy into a single entity that will sit lower in the management hierarchy. The agency will also eliminate the National Center for Environmental Research, which oversees studies on how people respond to chemical exposures, and transfer its workload to other offices. The moves will "reduce redundancies" and "streamline management," EPA said. But science advocates worry that the changes will make it harder for scientists to have a voice in decisions.

NIH funds sequencing centers

BIOMEDICINE | All of Us, the National Institutes of Health (NIH) precision medicine program that plans to enroll 1 million people, reached a milestone last week: NIH selected three genome sequencing centers to analyze participants' DNA. The Broad Institute in Cambridge, Massachusetts; Baylor College of Medicine in Houston, Texas; the University of Washington in Seattle; and partners will receive awards totaling $28.6 million the first year, to be extended up to 5 years. The centers will test participants' DNA samples for genetic markers and sequence their genomes. More than 110,000 people have registered for All of Us; of those, more than 60,000 have completed full enrollment, including filling out surveys and donating blood and urine.

A 'manel-free' cancer meeting

GENDER | Calling out "manels"—all male panels at meetings—is one way to raise awareness about the imbalance of men and women participating in scientific conference programs. Now, organizers of a meeting at a leading cancer research institute in Germany have gone a step further: At the Frontiers in Cancer Research meeting, hosted by the German Cancer Research Center in Heidelberg next week, 23 of the 28 invited speakers—82%—are women. Organizers considered inviting only women as speakers but decided instead to invite one man to speak at each session. They say they will be pleased if, at first glance, no one notices anything unusual about the program. So far the response has been positive, with nearly 250 participants registered.

Distant dwarf planet spotted

PLANETARY SCIENCE | A goblin lurks at the edge of our solar system. On 2 October, a team of astronomers announced the discovery of a new dwarf planet, provisionally named 2015 TG387, that follows the most distant orbit yet confirmed, traveling up to 2300 times farther from the sun than Earth does. Nicknamed "The Goblin" thanks to its discovery close to Halloween in 2015 by the Japanese Subaru 8-meter telescope in Hawaii, the 300-kilometer-wide planet takes 40,000 years to complete an orbit. For 99% of this time, it would have been too faint to see, but the astronomers, led by Scott Sheppard of the Carnegie Institution for Science in Washington, D.C., lucked out, discovering the planet near its closest approach to the sun. The Goblin now joins a small group of extreme solar system objects whose orbits hint at tugs from a hypothesized, but not yet observed, massive planet, called Planet Nine or Planet X, hidden in the fringes of the solar system.

CLIMATE CHANGE

A climate gap in fuel prices

The taxes that people in 42 developed countries pay to use oil, coal, and gas don't cover the costs of the climate warming they cause, the Organisation for Economic Co-operation and Development said in a report last week. The biennial report charted a carbon pricing gap: the extent to which surcharges on fossil fuels from taxes and emission permits fall short of matching the costs of storms, droughts, and other effects of climate warming. Across the 42 countries studied, the shortfall averaged 76.5% of climate warming costs. The report urges countries to increase taxes on carbon use to discourage consumption and make alternative energy sources more cost competitive. Countries are instituting new policies, it says, but the pricing gap is declining by just 1% per year—too slowly to meet the Paris agreement's targets for global emission reductions. (In this chart, 2015 is the most recent year for which country-level data are available.)

Percentage of climate warming costs covered by carbon taxes

Country	Percentage
Switzerland	73%
Norway	66
United Kingdom	58
South Korea	57
Italy	54
Spain	49
Germany	47
Canada	35
Poland	33
Japan	31
United States	25
Australia	21
India	14
China	10
Brazil	6
Indonesia	5
Russia	0

Countries with higher taxes on carbon use tended to rely less on fossil fuels.
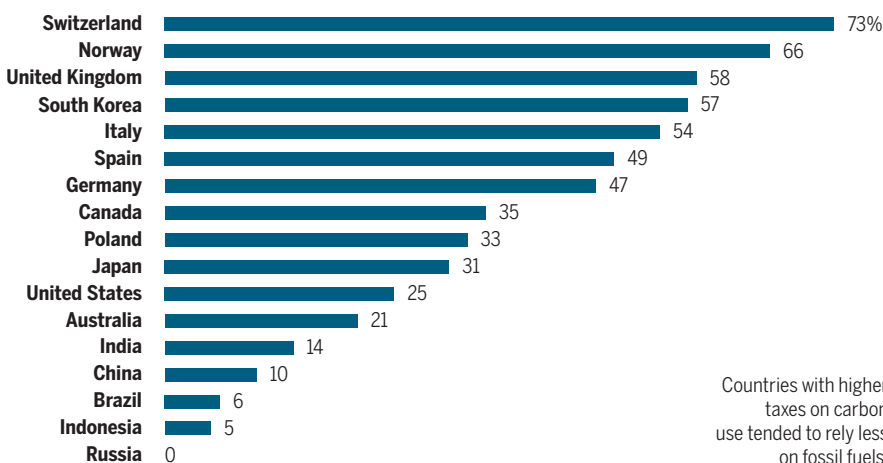
S SCIENCEMAG.ORG/NEWS
Read more news from *Science* online.

Mercury's northern hemisphere is covered by lavas (tan) in this false-color image.

PLANETARY SCIENCE

BepiColombo set to probe Mercury's mysteries

European-Japanese mission will examine idea that innermost planet formed out past Mars

By **Daniel Clery** *and* **Dennis Normile**

Tiny and relatively ignored, Mercury holds outsize mysteries. Only two spacecraft have made the difficult journey to its sunbaked environs. Now comes the planet's third and most ambitious visitor, the European-Japanese mission BepiColombo, a pair of probes due to launch on 20 October. Picking up where the last visitor, NASA's MESSEN-GER mission, left off in 2015, BepiColombo will probe puzzles including Mercury's skewed magnetic field, its overstuffed iron core, and strange lakelike depressions perhaps carved by escaping volatile elements.

"MESSENGER really threw into question many theories about how this planet came to be," says BepiColombo team member Emma Bunce of the University of Leicester in the United Kingdom. BepiColombo "is perfectly timed and set up to answer these questions," says mission scientist Johannes Benkhoff of the European Space Agency's (ESA's) technology center in Noordwijk, the Netherlands.

The first Mercury probe, NASA's Mariner 10, made a series of flybys in 1974 and 1975, some 40 years before MESSENGER. But getting a spacecraft into orbit around Mercury without it plummeting into the sun was a tough problem. In the 1980s, mission planners worked out the complex series of gravity assists from other planets needed for the journey, building on work by Italian astrophysicist Giuseppe "Bepi" Colombo, now honored by the new mission. A second challenge—how to endure solar radiation 10 times as strong as at Earth while observing a surface heated to 400°C—also delayed a second mission. "Spacecraft had never spent so long close to the sun," says Sean Solomon, director of the Lamont-Doherty Earth Observatory in Palisades, New York, who led the MESSENGER mission.

The ESA and NASA missions were both approved in the late 1990s, but the smaller $450 million MESSENGER mission got to the launchpad first and reached Mercury in 2011. The more ambitious €1.65 billion

The European Space Agency's orbiter will be joined by a smaller Japanese spacecraft.

BepiColombo hit a few bumps during development, including solar arrays that degraded too quickly, which delayed its launch about 5 years. Japan joined the mission in 2003, eager to build on its success observing a planetary magnetic field in the 1992 Geotail mission, which studied the tail of Earth's magnetosphere. "Mars and Venus don't have magnetic fields, so the target was Mercury," says Go Murakami, project scientist for Japan's part of BepiColombo at the Japan Aerospace Exploration Agency's Institute of Space and Astronautical Science (ISAS) in Sagamihara.

ISAS's contribution is a separate spacecraft with five instruments that focus on the magnetosphere, whereas the larger ESA orbiter, with 11 instruments, is equipped to study the planet itself. Following launch from ESA's spaceport in French Guiana, the two stacked craft will embark on a 7-year journey, swinging once past Earth, twice past Venus, and six times past Mercury before finally separating and entering orbit around the planet in 2025.

One mystery awaiting the mission is MESSENGER's discovery of many volatile elements on the planet's surface, including chlorine, sulfur, potassium, and sodium, which should have been boiled off by the sun's heat long ago. "There is something odd in the formation history of Mercury," Benkhoff says. A clue comes from the ratio

of potassium to thorium, which indicates a planet's temperature during formation. Benkhoff says Mercury's ratio points to a cooler origin, farther out than Mars. Volatiles are more abundant at those distances, and if Mercury formed beyond Mars and drifted in only later, it would have retained a larger supply of volatiles.

BepiColombo will make a sharper map of the volatiles than MESSENGER, thanks to an imaging spectrometer that identifies elements by how they fluoresce when hit by solar x-rays. That could help it track the loss of volatiles today. Depressions spotted by MESSENGER, tens of meters deep and hundreds wide, could be formed by escaping gases. If any have changed since MESSENGER's visit, it will suggest a large volatile supply that continues to vent into space.

BepiColombo will also scrutinize the peculiar contrasts between the planet's north and south. The north has large areas covered by smooth volcanic material that must have erupted relatively recently, whereas the south is cratered and ancient. Mercury's magnetic field mirrors the divide, as it, too, is shifted to the north. "How are these asymmetries connected?" Solomon asks.

Inside Mercury is another anomaly for the mission to explore: the planet's huge iron core, extending out to 80% of its radius, surrounded by a relatively thin layer of rock. One theory is that early in life, Mercury suffered a collision with another body that blasted off most of the lighter rocky material, leaving the heavier iron behind. Researchers would expect the iron core to have cooled and solidified by now, yet at least some of it is still liquid and churning, generating a magnetic field.

Although the field is 100 times weaker than Earth's, it accelerates electrons from the solar wind to high energy levels, a phenomenon not seen in Earth's magnetosphere. Magnetometers aboard both the European and Japanese spacecraft should help the team understand the processes behind the energy boost, says James Slavin, a lead MESSENGER investigator at the University of Michigan in Ann Arbor, who expects "definitive answers to the mystery."

Studies of the magnetosphere could have implications beyond the solar system. Exoplanets found orbiting cool red dwarf stars could host liquid water—and conceivably life. But because they orbit their stars more closely than Mercury orbits the sun, they likely face strong stellar winds and radiation levels inimical to life—unless the planet is protected by a magnetosphere like Mercury's. "If we want to understand if life can [survive] on such planets, one of the important bits of information is the magnetosphere," Murakami says. ∎

Trio earns physics Nobel for turning lasers into tools

Prize honors discoveries that led to optical tweezers and short, powerful bursts of laser light

By **Adrian Cho**

If you've had laser eye surgery and chucked your clunky glasses, you can probably thank some of the winners of this year's Nobel Prize in Physics. The award, announced on 2 October, honors a trio of physicists who invented laser tools that have become ubiquitous in research and beyond. The winners include the oldest person to win a Nobel and the first woman to win the physics prize in 55 years.

Arthur Ashkin, 96, received half of the $1 million prize for the invention of optical tweezers, a technique that uses focused laser beams to hold and manipulate tiny objects, including viruses, bacteria, and individual cells. Recognition of the advance, which Ashkin made in 1986 at the famed Bell Laboratories in Holmdel, New Jersey, "is well overdue," says Philip Jones, an optical physicist at University College London.

The other half of the prize honors Gérard Mourou, 74, of the École Polytechnique in Palaiseau, France, and Donna Strickland, 59, of the University of Waterloo in Canada, for their invention of chirped pulse amplification (CPA), a technique that boosts the intensity of ultrashort laser pulses. Strickland is the first woman to win the physics Nobel since Maria Goeppert Mayer in 1963. "Finally," says Anne L'Huillier, an atomic physicist at Lund University in Sweden.

Optical tweezers use the subtle forces of light to tug on tiny objects. The electric field in laser light can polarize a transparent material, such as plastic or cellular cytoplasm. The polarized material then experiences a force drawing it to where the electric field is strongest, typically the middle of the beam. Move the beam, and you can pull on the object.

Ashkin did his original work with microscopic plastic beads, which can be latched onto the end of a DNA molecule to stretch it out. But optical tweezers can also pluck and prod living cells themselves. Caroline Adiels, a biologist at the University of Gothenburg in Sweden, employs them to measure the mechanical properties of yeast,

bacteria, and diatoms. "I use the technique on a daily basis," she says.

Mourou and Strickland invented CPA when physicists were struggling to boost the intensity of very short laser pulses. Once a laser pulse grows too intense, it distorts the properties of the amplifying medium, preventing further amplification, explains John Collier, a laser physicist and director of the Central Laser Facility at Rutherford Appleton Laboratory in Didcot, U.K.

Strickland was Mourou's graduate student at the University of Rochester in New York when, in 1985, the two found a clever way around this problem. Their scheme first splits a laser pulse into its frequency components by, say, reflecting it off a diffraction grating. Those spread-out components can then be run through separate amplifiers without overburdening them. Finally, the beefed-up components are recombined into a short pulse far more intense than the original.

CPA is now the basis for most ultrafast laser systems—including those used in laser eye surgery, in which a pulse of laser light quickly slices open the cornea before the surrounding tissue has time to heat up. "I'm glad the committee has recognized this science, which is creating so much value," says Toshiki Tajima, a laser physicist at the University of California, Irvine. By packing energy into such a short pulse, the lasers can also achieve petawatts of power, useful for creating high temperatures, powerful electromagnetic fields, and other "extreme conditions that you wouldn't otherwise find on Earth," Collier says.

Both L'Huillier and Adiels praise the award for recognizing the contributions of women in physics. "Tiny steps, taking forever, but at least something is happening," Adiels says. Whereas 206 men have won or shared the physics prize, only three women—including Strickland—have received the honor. That skewed tally appeared to surprise her at a press conference announcing the prize. "Is that all?" Strickland asked. "Really?" ∎

With reporting by Daniel Clery.

PHYSICS NOBEL

"for groundbreaking inventions in the field of laser physics."

Arthur Ashkin
Gérard Mourou
Donna Strickland

Tasuku Honjo (left) and James Allison (right) won the Nobel Prize in Physiology or Medicine.

NOBEL PRIZE

Cancer immunotherapy sweeps Nobel for medicine

Prize goes to discoveries that unleash immune system

By **Jocelyn Kaiser** *and* **Jennifer Couzin-Frankel**

Fittingly, James Allison heard the news at a conference of the very field his work helped launch. At a New York City hotel where he was staying for a meeting on cancer immunotherapy, Allison, an immunologist at University of Texas MD Anderson Cancer Center in Houston, received a 5:30 a.m. phone call on 1 October from his son, then answered a knock at his door to find colleagues carrying champagne. That day, he and Tasuku Honjo of Kyoto University in Japan were awarded the Nobel Prize in Physiology or Medicine for work underpinning a new class of drugs called immune checkpoint inhibitors, which have revolutionized the treatment of certain types of cancer.

In basic research beginning in the 1990s, Allison and Honjo discovered ways to release the immune system's "brakes," which prevent it from attacking tumor cells. Neither set out to conquer cancer; both were doing fundamental studies of the immune system. But the treatments resulting from their work are now causing previously untreatable tumors, such as advanced melanoma, to disappear for years in some patients. "The time is right," Klas Kärre, an immunologist at the Karolinska Institute in Stockholm and a member of the Nobel Committee, said at a press conference. "The first approved drug based on this treatment came in 2011. … We can now see the long-term outcome. It's very convincing."

MEDICINE NOBEL

"for their discovery of cancer therapy by inhibition of negative immune regulation."

James Allison
Tasuku Honjo

Both awardees shared how thrilled they were that their basic research had led to lifesaving therapies. "More than anything, I'm happy to realize that my research is so meaningful" for patients, Honjo, 76, said after the prize was announced. Allison, 70, told reporters: "To have my work really impact people … it's everybody's dream."

Researchers began to think that activating the immune system could fight tumors more than 100 years ago. But efforts to develop treatments based on this idea—with vaccines, for example—foundered in part because researchers didn't understand enough about how the immune system works.

Allison built on the work of French immunologists in the 1980s who were studying T cells, sentries of the immune system that attack pathogens and other foreign invaders. They identified a key receptor on the surface of T cells that they called cytotoxic T-lymphocyte antigen 4, or CTLA-4. Allison and others found that the receptor puts the brakes on T cells, preventing them from launching fullout immune attacks.

Other groups hoped to use the receptor to help treat autoimmune diseases, in which the immune system's brakes aren't strong enough. But Allison had a different idea. He wondered whether blocking the blocker— the CTLA-4 molecule—would free the immune system to destroy cancer. In a paper in *Science* in 1996, Allison reported that antibodies against CTLA-4 erased tumors in mice.

Pharmaceutical companies initially shied away from developing the therapy, wary of possible side effects and of an approach so radically different from the standard treatments of surgery, radiation, or chemotherapy. A small biotechnology company, Medarex, in Princeton, New Jersey, acquired rights to the antibody in 1999. In a series of trials, the drug caused tumors to shrink—sometimes for a decade or more—in about 20% of patients with melanoma that had spread.

Honjo followed a similar track. In 1992, he discovered a molecule expressed in dying T cells, which he called programmed death 1, or PD-1. He recognized PD-1 as another brake on T cells. Later he and others showed that blocking PD-1 caused tumors to shrink in mice. PD-1 blocking drugs are now approved for lung, kidney, and bladder cancers, along with melanoma and others. Clinical trials in other tumor types are underway.

The Nobel is "great validation for our field. And certainly Jim and Tasuku were real pioneers and two of the key leaders," says Drew Pardoll of Johns Hopkins University in Baltimore, Maryland. But Pardoll notes that other researchers contributed key early discoveries. A Nobel in the field was widely expected, but Pardoll says "it's always a little bit challenging" to predict actual winners.

To other immunotherapy researchers, this year's medicine award is especially gratifying; until relatively recently the field was considered long on promise but short on data. "We were in the last paragraph of the book, the last session of the meeting on the last day," says pediatric oncologist Crystal Mackall of Stanford University in Palo Alto, California.

No longer. The Fourth International Cancer Immunotherapy Conference in New York City lasted 4 days and had more than 1000 participants, who gave Allison a standing ovation. ∎

With reporting by Dennis Normile and Gretchen Vogel.

MEXICO

New science minister's activism sparks debate

Biologist Elena Álvarez-Buylla has spoken out against transgenic maize

*By **Lizzie Wade**, in Mexico City*

In early June, evolutionary developmental biologist Elena Álvarez-Buylla received an out-of-the-blue phone call from the campaign of Andrés Manuel López Obrador, then the front-runner in Mexico's presidential election, with a question. If López Obrador won, would she consider becoming the next director of the National Council of Science and Technology (Conacyt), the country's science ministry and primary granting agency? "My first reaction was to say, 'I can't,'" recalls Álvarez-Buylla, a professor at the National Autonomous University of Mexico (UNAM) here. "I have a great passion for scientific research," and she couldn't imagine leaving the laboratory.

But after thinking it over for a few hours, her passion for public service took over. "I started to have a feeling that I couldn't say no," says Álvarez-Buylla, who founded and leads Mexico's Union of Scientists Committed to Society (UCCS). "It doesn't matter how big the personal sacrifice is. … This is a unique and historic moment" for Mexico.

López Obrador, a progressive populist, won the presidency in a landslide and will be sworn in on 1 December; Álvarez-Buylla is now preparing to leave the lab bench and assume her new role. She will be the president's primary science adviser and determine priorities for Conacyt's approximately $1.5 billion budget, which funds grants to scientists working in the public and private sectors and supports tens of thousands of Mexican students at home and abroad.

Many scientists are delighted that one of their own will lead Conacyt—most of Álvarez-Buylla's predecessors were career administrators—and that she'll be the first woman to do so. But critics worry about her opposition to genetically modified (GM) maize, which Álvarez-Buylla fears could spoil the country's astonishing agricultural biodiversity. They also worry that in her commitment to socially relevant science, she may neglect basic research. A petition asking López Obrador to pick another director has gathered more than 1000 signatures.

"There's not a clear boundary" between her research and her activism, says Rodrigo Álvarez Aguilera, a science teacher here and one of the petition's organizers. Biochemist Luis Herrera Estrella, director of the National Laboratory of Genomics for Biodiversity in Irapuato, says Álvarez-Buylla is "a very good scientist" but calls her views on GM organisms "radical."

Born into a family of scientists, Álvarez-Buylla studied plant biology at UNAM and received her doctorate from the University of California, Berkeley. She returned to UNAM in 1992, where she now runs several research groups. Colleagues praise her contributions to the understanding of plant root development and how plant genotypes influence their traits. "There's no question that the research she does is fantastic," says her former collaborator Chelsea Specht, a plant evolutionary biologist at Cornell University. "And her advocacy is based in very good research."

That advocacy began after a 2001 *Nature* paper reported that genetic material from the cauliflower mosaic virus, a common addition to GM plants, had been found in native maize varieties sampled in Mexico's southern state of Oaxaca—likely the result of cross-pollination from industrially grown crops whose origin remains unclear. The finding shocked many because of maize's all-important role in Mexican history and culture. Maize domestication began here about 9000 years ago, and Mexico now boasts at least 59 native varieties, called landraces, each exquisitely adapted to regional environmental and climatic conditions. Some possess unusual characteristics; in August, for instance, researchers reported that a landrace from Oaxaca can fix nitrogen from the atmosphere with the help of microbes, a trait known from beans and other legumes but never before found in maize. The adaptation allows it to thrive in nitrogen-poor soils; breeding it into other maize varieties could be a boon for farmers and might help reduce fertilizer use. Other landraces may have useful adaptations as well.

Álvarez-Buylla led a team that confirmed the results of the 2001 study and has continued to hunt for transgenic DNA and any possible effects in Mexican landraces, work that helped her win Mexico's National Science Prize in 2017. She says she has nothing against genetic engineering in itself;

Corrected 4 October 2018. See full text.

"I'm not a Luddite who is scared of technology," Elena Álvarez-Buylla says.

her team creates and studies GM plants in the lab, and such experiments should not be prohibited or restricted, she says. "I'm not a Luddite who is scared of technology." But her own experiments have shown introduced genes can have unpredictable effects. "If a transgene is inserted in one part of [a plant's] genome, it can be silenced and have no effect. If it's inserted in another part, it can lead to a tremendous change," she says. That unpredictability makes it too risky to allow GM maize anywhere near Mexico's landraces, she argues. Planting GM maize in Mexico has been prohibited since 2013, pending the outcome of a lawsuit. Álvarez-Buylla has been an outspoken proponent of a permanent ban.

Herrera Estrella, who develops GM plants, disputes the risks. In the more than 50 years that landraces have cross-pollinated with commercial, industrially grown maize varieties, "it has had no negative effect. … It's not contamination. It's a completely natural biological process," he says. If a foreign gene harms the growth or development of maize, farmers simply won't use seeds from that individual plant the next year, and the damage won't be passed on. "Conacyt needs a director with an expansive view of science and technology and the impact they could have on Mexico's development. I think she has some very biased opinions," he says.

Some scientists have also raised questions about Álvarez-Buylla's plans to open dialogue between Mexico's scientific community and Indigenous knowledge producers. "Mexico has the opportunity to contribute to the world something truly new that comes out of a deep hybridization" of those two forms of knowledge, Álvarez-Buylla says. Critics say that philosophy diminishes Western scientific values and achievements. They also worry basic research will suffer because of Álvarez-Buylla's professed commitment to science that helps solve societal problems such as infant mortality and dwindling water supplies. "That's a fallacy," she says; Conacyt will remain supportive of basic research in all fields.

Álvarez-Buylla has little administrative experience, but she has proved herself capable at building and managing large research groups, says longtime colleague Daniel Piñero Dalmau, a plant population geneticist at UNAM. She has promised to improve Conacyt's grant evaluation process, which many say is frustratingly quantitative and completely opaque. "Changing anything in Mexico is really hard," Piñero Dalmau says. But Álvarez-Buylla "is deeply committed to this country." ■

ANIMAL RESEARCH

Airlines fight effort to force them to carry lab animals

U.S. Department of Transportation weighs whether to investigate complaint by leading research organization

By **David Grimm**

A last-ditch attempt by biomedical science advocates to force airlines to transport nonhuman primates and other research animals appears to be facing stiff headwinds. Last week, four international carriers strongly urged the U.S. Department of Transportation (DOT) to summarily reject a plea from a leading research advocacy organization to order the airlines to resume flying study animals to research facilities around the world. The request is "misguided," "far-fetched," and contrary to laws that allow airlines to decide what kinds of cargo they will carry, the companies argued. DOT has not said how it will respond.

The forceful language came in response to a complaint sent to DOT this past August by the National Association for Biomedical Research (NABR) in Washington, D.C., which represents universities, lab animal providers, and other groups. NABR alleges that the four carriers "illegally discriminate" against customers who transport research animals because the airlines willingly transport the same animals as pets and for zoos. The bans "will slow down the progress of essential and life-saving biomedical research," says NABR President Matthew Bailey.

NABR wants DOT to investigate United Airlines, British Airways, China Southern Airlines, and Qatar Airways. And it wants the agency to fine any airline—including more than two dozen other major carriers—that refuses to reverse its ban.

The filing is "an act of desperation," says Kathy Guillermo, a senior vice president at Norfolk, Virginia–based People for the Ethical Treatment of Animals (PETA), which has targeted airlines that ship research animals for nearly a decade.

Such efforts appear to be working. In 2013, United Airlines stopped carrying research animals, saying it was responding to pressure from animal rights groups and concerns about passenger safety. In July, the Russian carrier AirBridgeCargo became the latest carrier to announce a ban after receiving 200,000 emails generated by a PETA campaign. The moves mean almost all major airlines now don't transport nonhuman primates, and most bar any animals bound for scientific laboratories. A major exception is Air France, which says it wants to support biomedical science.

More than 1700 projects funded by the U.S. National Institutes of Health rely on nonhuman primates, according to NABR, which notes that the U.S. Food and Drug Administration often requires researchers to use these animals when testing human drugs.

To get around the bans, many labs and firms in the United States now truck nonhuman primates around the nation. Some use chartered aircraft to import animals from abroad. In Spain's Canary Islands, researchers were forced to use military planes to get lab mice. And last year, the island nation of Mauritius began to invite researchers to come there to study its large population of macaques so that they wouldn't have to deal with shipping problems.

Such practices can complicate studies and drive up costs. Some researchers "are being priced out of being able to do this work," or moving studies to nations "where they don't have the same level of animal welfare regulation," says Cindy Buckmaster, chair of the board of directors of Americans for Medical Progress, a nonprofit in Washington, D.C., that promotes the need for animals in labs.

NABR filed a similar, but informal, complaint with DOT last year, and the agency took no action. Both sides are planning to file additional arguments with the agency. Whatever happens, Buckmaster says, "The need for research animals doesn't end just because people don't want to fly them." ■

> *The airline bans "slow down the progress of essential and life-saving biomedical research."*
>
> **Matthew Bailey,** National Association for Biomedical Research

Corrected 4 October 2018. See full text.
Published by AAAS

Streams of stars surround the galaxy NGC 5907, relics of a shredded dwarf galaxy. Dozens of streams have recently been found around the Milky Way.
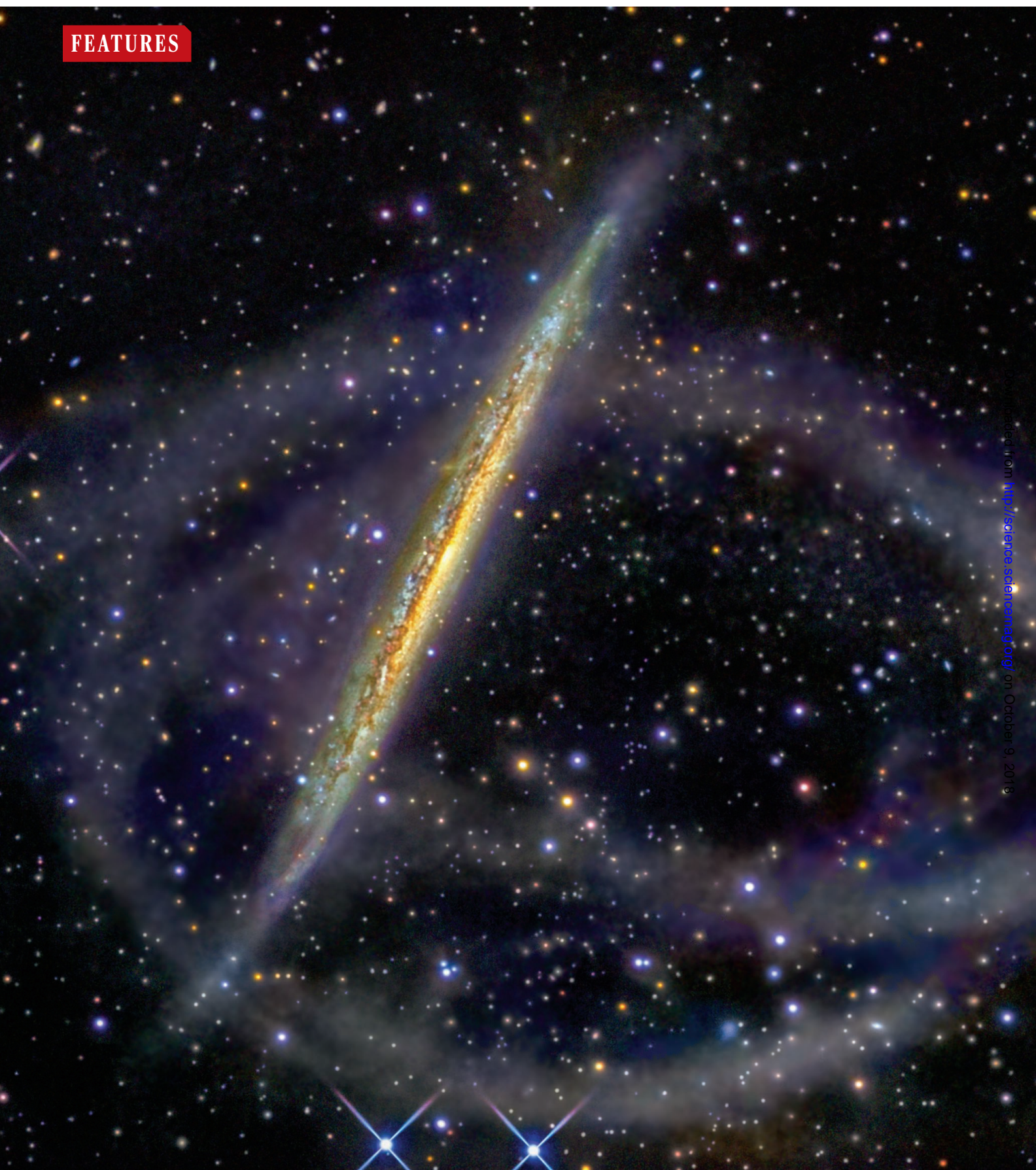
SKY RIVERS

Streams of stars falling into our galaxy trace its history and mass. They may even record encounters with clumps of dark matter

By **Eric Hand**

At 4 a.m. on Zero Day, Ana Bonaca woke up in her Manhattan hotel in a state of excitement. She made her way through the hushed streets of New York City with several other early bird astronomers. "It was still dark, and raining," says Bonaca, a Harvard University postdoc. At the Flatiron Institute, a new and lavishly funded center for computational science, they settled into a conference room on the empty third floor. The hunt for galactic ghosts—witnesses to the Milky Way's violent history—was about to begin.

Zero Day was the name for an impending data release from Gaia, a European Space Agency satellite mapping the Milky Way. At 6 a.m. on 25 April, the mission was to release positions and motions for more than a billion stars, all at once, as a model of open science and a way to prod publication-hungry astronomers into action. "It was a radical way to release the data," says Adrian Price-Whelan, a Princeton University postdoc who also arrived early at the Flatiron Institute. "Everyone felt pressure." By day's end, nearly 70 astronomers would converge there, some from as far away as Australia, to sift through the data together.

Many planned to study the stars of the Milky Way's whirling disk. But Bonaca and Price-Whelan's interests were farther afield. They would be examining the galaxy's halo, a sparse and roughly spherical region of stars that envelops the disk like a snow globe. Stretching across that firmament, they knew, were stellar streams: the filamentous remains of neighboring small galaxies and star clusters that had been disemboweled as they fell into the Milky Way's gravitational grasp. "They're almost like necklaces of stars," Bonaca says.

With astronomical surveys like Gaia pushing into the far reaches of the Milky Way's halo, the study of stellar streams is entering a golden age. Even before Gaia, astronomers were finding streams in data from ground-based observing campaigns such as the Sloan Digital Sky Survey and the Dark Energy Survey. "There has been an explosion in the number of streams that has been accelerating as more surveys come out," says Kathryn Johnston, a galactic dynamicist at Columbia University. In just the past 2 years, the count has more than doubled, passing 60.

It is more than mere stamp collecting, says Vasily Belokurov, an astronomer at the University of Cambridge in the United Kingdom. "Before, we were looking at them in awe," he says. "Now, we are starting to use them as tools." The streams are particularly useful for what astronomers call galactic archaeology—rewinding the cosmic clock to reconstruct the assembly of the Milky Way. They also are being used as exquisitely sensitive scales to measure the galaxy's mass.

The third possible application for stellar streams is perhaps the most intriguing. Patrolling the outskirts of the galaxy's halo, the streams are well-positioned to reveal the presence of dark matter, the unseen stuff thought to dominate ordinary matter by a ratio of nearly six to one. Because the streams are so fragile, theorists say, collisions with marauding clumps of dark matter could leave telltale scars, potential clues to its nature.

One stream in particular interested Bonaca and Price-Whelan, who knew each other but had never collaborated. As Zero Day began, David Hogg, an astronomer at New York University in New York City

Gravity's rainbows

The Milky Way's gravity stretches satellite galaxies and star clusters into long, slender streams. They show how galaxies grow by consuming their neighbors, and can be used to measure our galaxy's mass. Furthermore, gaps in the streams could mark the sites of encounters with dark matter, the unseen stuff that constitutes 85% of the universe's mass.

Palomar 5 (Pal5) stream
Pal5 was discovered in 2001 as the first thin stream formed from a globular cluster. Its current orbit takes it far over the galactic center.

Globular clusters
These hives typically hold 100,000 stars or fewer and give rise to long, thin streams.

Gap

GD1 stream
Discovered in 2006, GD1 is the longest known thin stream, stretching across more than half the northern sky. It contains a gap that could be the scar of a dark matter collision 500 million years ago.

Sun

Milky Way

A halo of darkness

The Milky Way resides in a vast halo of dark matter that could be anywhere from semi-spherical to pickle shaped. Star streams could yield clues to its shape, and whether it contains subhalos—denser clumps of dark matter.

Milky Way — Dark matter

Subhalo

1.3 million light-years
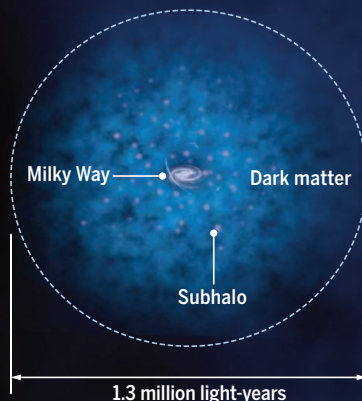
A surge of streams

New sky surveys have accelerated discoveries of distant streams.

20

10

0

1990   2000   2010   2020

1

2

3
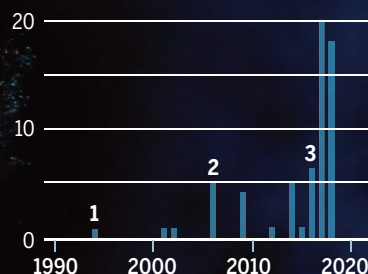
1 Hints of the first stream, Sagittarius, discovered.

2 Data from the Sloan Digital Sky Survey spur discoveries.

3 Gaia, the Dark Energy Survey, and Pan-STARRS drive a new surge.

Dwarf galaxy
With millions of stars, dwarf galaxies form fatter streams.

Drawn and quartered

Streams are created by the tidal forces of the Milky Way's gravity. Gaps in streams could point to dark matter encounters.

Stream formation

1 A dwarf galaxy or globular cluster strays too close to the galaxy.

2 Tidal tails form because of differences in the pull of the Milky Way on opposite sides.

3 The leading and trailing tails rotate to parallel the satellite galaxy's orbit.

Orbit direction

Tidal tail

Galaxy center
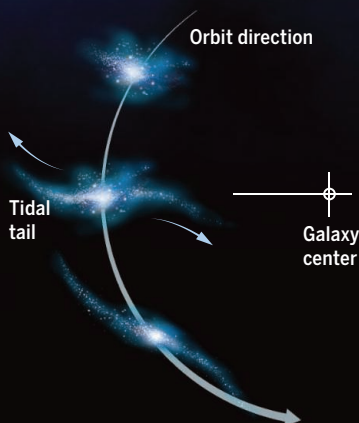
Gap formation

1 A mature stream orbits around the Milky Way.

2 A dark matter subhalo crosses the stream, causing fishhook-shaped spurs to form.

3 The spurs settle back into the stream, leaving a gap between thickened edges.

Orbit direction

Subhalo encounter

Gap
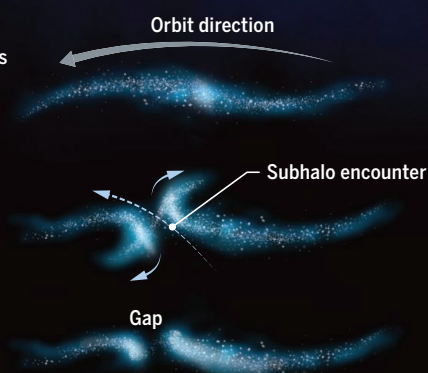
who helped organize the Flatiron workshop, suggested they see what Gaia had to say about GD1, one of the longest and most slender streams in the sky. GD1 had been studied since its discovery in 2006, but Hogg wondered whether there might be surprises in the sharper picture that Gaia promised. "We didn't appreciate the scope of his intuition until much later," Price-Whelan says.

THE ACCEPTED STORY of our galaxy goes like this: Soon after the big bang 13.8 billion years ago, dark matter coalesced under its own gravity into a cosmic web of clumps and filaments. Ordinary matter—clouds of hydrogen gas—settled into those gravitational wells and ignited to form the first stars, nestled in the first, small galaxies. Bigger galaxies like our own grew by gobbling up their small neighbors.

Several hundred miniature galaxies and star clusters still orbit the Milky Way today, and astronomers have long looked for evidence that they are succumbing to our galaxy's gravity. The first sign would be "tidal tails"—growths on opposite sides of the satellite galaxies and clusters, analogous to the bulging ocean tides on opposite sides of Earth. On Earth, the ocean tide on the side facing the sun or moon is intuitive: Gravity is stronger there. But the tide bulges on the opposite side, too, where gravity is weaker, as the solid earth is tugged out from under the ocean. Replace the sun or moon with the Milky Way, and the same mechanism shapes the satellites.

In the 1990s, computer models showed the tails were just the beginning. Under the Milky Way's influence, stars would unspool from both sides of a satellite until the thread ran out. Crucially, the models showed that the streams could persist for billions of years, through multiple galactic orbits, before being shredded into anonymity. Therefore, many of the Milky Way's recent acquisitions should still be visible. "My generation was the one that realized we should start seeing them around our own galaxy," Johnston says. "And then indeed we did."

Astronomers first saw star streams veiling the fringes of nearby large galaxies. Discerning those tenuous smears of stars around our own galaxy is far harder, given Earth's vantage inside the Milky Way's star-rich disk. Yet in 1994, astronomers stumbled on the first homegrown example: Sagittarius, named after the constellation in which it was discovered. It was a lumpy dwarf galaxy in the final throes of being devoured, hanging off the bottom of the

Milky Way. At first, researchers saw mere hints of it. But in 2003, an infrared survey sensitive to the rare, red giant stars in the stream showed how it wraps around the Milky Way like a claw.

The spectacular find was a vindication for what astronomers call the hierarchical theory of galaxy formation: Here was the voracious Milky Way, caught in the act of eating a smaller cousin. But Sagittarius wasn't a stellar stream so much as an Amazonian river of stars, too mixed up with the Milky Way to be much use for gauging the galaxy's mass and makeup.

Better would be the thinner, more distant streams expected to form from globular clusters, the second kind of Milky Way satellite. Whereas dwarf galaxies hold millions of stars, globular clusters typically contain fewer than 100,000. The Sloan survey, a catalog of millions of stars and galaxies, was providing just the sort of deep, all-sky data that might reveal such

> *"Before, we were looking at them in awe. Now, we are starting to use them as tools."*
>
> **Vasily Belokurov**, University of Cambridge

slender streams. Almost overnight in 2001, astronomers discovered Palomar 5 (Pal5): a long, thin stream hovering above the galactic center. Carl Grillmair, an astronomer at the California Institute of Technology (Caltech) in Pasadena, soon found that Pal5 stretched a full 24° across the sky.

Then he found an even longer stream. Grillmair was perfecting a technique to distinguish distant stream stars from crowds of foreground stars. He knew dwarf galaxies and globular clusters, the wellsprings for streams, are ancient and metal-poor; as a result, their stars tend to be bluer than average. So he examined slices of the sky at ever fainter magnitudes—the equivalent of looking ever farther out into the halo—and filtered by color. By stacking the slices and moving through them like a flip book, he could see when a population of bluish stars jumped out. "I'd watch these movies—out, back, out, back—through the galactic halo."

One weekend in late 2005, soon after the Sloan survey released fresh data, Grillmair remembers scrolling through a virgin part of the sky on his home computer. His wife looked over his shoulder. "Do you see that?" he recalls asking her. "Is that real?"

It was real: an impossibly thin stream stretching across 60°, or a third of the sky's hemisphere. It was about 26,000 light-years away from Earth, far removed from the galactic center and its convulsive effects.

THE GOSSAMER STREAM, called GD1 after Grillmair and his Caltech collaborator Odysseas Dionatos, was the first of more than a dozen streams Grillmair later discovered—more than anyone else. But its distance and seeming fragility have made it a favored tool for astronomers trying to weigh the galaxy. Crude mass readings can be gleaned from stars or gas in the disk, which whirl around the galactic center: The faster they orbit, the greater the mass tugging them inward must be. "This is how [Johannes] Kepler and [Isaac] Newton figured out gravitational forces in the solar system," says Jo Bovy, an astrophysicist at the University of Toronto in Canada. But stars in the roiling disk rarely move in perfect circles, which throws off the measurements. And the technique tells you only about the mass of the disk, whereas most of the Milky Way's mass resides in its large dark matter halo.

Intact dwarf galaxies and globular clusters lie farther out and are sensitive to the halo mass, but their orbital paths are unknown. Astronomers use their average motions to make mass estimates, but the results vary wildly, between 700 billion times the mass of the sun and several times that much. Astronomers aren't sure whether the Milky Way weighs more or less than Andromeda, our nearest major galaxy.

In contrast, the shape of a stream clearly indicates how individual stars are moving. "You have lots of test particles lined up for you on the sky," Belokurov says. To measure the enclosed mass out to the orbit of the stream, astronomers need only know the stream's shape and the speed of the stars along its path. The farther out and longer the stream, the better. In 2016, Bovy used Pal5 and GD1 to calculate the Milky Way's total mass: 800 billion solar masses, on the skinnier side of the estimates and much less than Andromeda.

The streams also let Bovy work out the shape of the inner part of the dark matter halo. An earlier analysis based on the Sagittarius stream had suggested a hockey puck, perpendicular to the disk, but Bovy's result was reassuringly conventional: a slightly flattened sphere. As astronomers begin to discover streams even more distant than GD1 and Pal5, they should be able to trace the outer part of the halo, which could take on stranger shapes, like that of a squashed pickle. "That's the dream for all of us," says Denis Erkal, an astrophysicist at the University of Surrey in the United Kingdom.

Other streams could map mass in finer detail. The Tucana III stream, for instance, discovered in data from the Dark Energy Survey, appears to have had a close en-

counter with the Large Magellanic Cloud—the Milky Way's largest satellite galaxy—several hundred million years ago. Erkal is hoping the imprint left on Tucana III by that encounter could be used to weigh that satellite. Similarly, the Ophiuchus stream passes close to the Milky Way's bar, a linear feature in the galaxy's central bulge. Distortions in the stream might trace the mass of the bar and how fast it rotates.

AT THE FLATIRON, Price-Whelan was hoping to map GD1 in finer detail than ever before. By 8 a.m. on Zero Day, fortified by free food and coffee, the astronomers were engrossed in getting Gaia data onto their laptops. Price-Whelan asked for the 80°-by-10° swath of sky that includes GD1. He slimmed the download by trying to exclude foreground stars that polluted the view of GD1. By 10 a.m., he had a streamlined, 1-gigabyte file, which he shared with Bonaca on a thumb drive.

The Gaia information was alluring to the two researchers because it contained the stars' motion across the sky. Most stellar streams had been discovered through some variation of Grillmair's filtering scheme, which offers a snapshot of a stream's probable members. But the stars in a stream also flow along paths that reflect the orbit of the disrupted satellite galaxy or star cluster. Gaia was the first sky-mapping effort to measure those "proper motions" for distant stars all at once. It was a whole new way of seeing streams.

Proper motions were especially powerful for picking out GD1 stars, Price-Whelan knew, because the stream orbits the Milky Way counter to the whirling stars of the disk. By sorting for stars with radically different motions from the rest, he cut the more than 2 million stars he had in his sky swath down to about 70,000 that were more likely to be a part of GD1.

While Price-Whelan fine-tuned his velocity cuts, Bonaca used another technique to further sharpen the picture. She set out to cross-match his probable GD1 stars with her own set, which she had sifted on the basis of color and faintness from data gathered by Pan-STARRS, a ground-based survey in Hawaii. In her hotel room after the second day of the workshop, she finally succeeded and made a quick plot. It revealed 1300 stars in GD1, which she and Price-Whelan would soon trace across 100° of the sky.

The detailed picture of the stream also showed something more. "OMG," she wrote when she emailed it to Price-Whelan. Next, she sent the picture to her former graduate school adviser. "My turn not to sleep," she wrote.

When they returned for the third and final day of the Flatiron workshop, Bonaca and Price-Whelan found a quiet glass-walled conference room. "Everyone sensed we were doing something that was not to be interrupted," Bonaca says.

The reason for their intensity was simple: They were seeing two gaps in GD1, as clear as day. One was the original location of the progenitor star cluster, its stars now drained away into the stream. The second gap was far more exciting. Though neither dared say it out loud, both were thinking: Here was the possible scar of a dark matter collision.

They showed the plot to Johnston, who had been Price-Whelan's graduate school adviser. She was stunned not only by the gap itself, but also by a peculiar structure on one side: a spur of stars, off the main track, as though the stream had folded back on itself. It reminded her of a feature her old simulations had predicted for a dark matter collision. "When I saw that, I was real excited,"

Johnston recalls. "I pulled up our old papers and said, 'Look, look, there it is!'"

Paranoid that other astronomers were pursuing the same trail, Bonaca and Price-Whelan didn't let up. After the workshop ended, they set up shop in coffee houses around New York City. "We wrote all weekend," Price-Whelan says. Around noon on 1 May—less than a week from the start of Zero Day—he posted their paper on the arXiv server and submitted it to *The Astrophysical Journal Letters*.

IN THE PAPER, published in August, Bonaca and Price-Whelan hint that GD1 was the victim of a galactic hit-and-run. In the prevailing models of galaxy formation, dark matter forms not only a giant semispherical orb—the halo itself—but also much smaller clumps of extra-dense dark matter that roam the larger halo. Some of those "subhalos" might be big enough to form the seeds of dwarf galaxies, but those weighing less than a billion solar masses would be too small to attract any ordinary matter. Starless, they would be invisible but for their gravitational distortions of other objects. Streams, being so long and fragile, would be especially likely to bear such scars.

Subhalos are not the only objects that can wound the streams. Giant clouds of gas liv-

ing near the disk also could plow through them. But the gap in GD1 was telling, not only because the stream's distant orbit probably puts it beyond the reach of the gas clouds, but also because of the gap's fine-grained structure—especially the spur that so captivated Johnston.

When a subhalo passes through a stream, stars both ahead of and behind the collision site should be gravitationally tugged toward the massive intruder. Stars ahead of it will lose rotational momentum and slip to a lower orbit around the Milky Way, where paradoxically they pick up speed and overtake other stream stars. Stars in the trailing edge gain rotational momentum and rise to a higher orbit, where they slow down and lag behind the stream. The result is fishhook-shaped spurs, visible on one or both sides of the gap, which eventually settle back into the stream (see graphic, p. 18).

To Erkal, the gap and spur suggest a dark matter subhalo collided with GD1. "I'd say it's very promising."

Others are not yet convinced. "Seeing a gap in one single stream is interesting—definitely interesting—but it's not incontrovertible evidence," says Alex Drlica-Wagner, an astronomer at Fermi National Accelerator Laboratory in Batavia, Illinois.

Confirmation would do more than just validate the existence of subhalos and the models of galaxy formation that predict them. By compiling statistics across many streams on the number and size of the gaps, astronomers might learn about the nature of dark matter itself. Warm, light dark matter particles would move too fast to clump into subhalos, leaving the streams unscarred by dark matter. The presence of gaps could imply colder, heavier dark matter particles, with smaller, more numerous gaps pointing to greater masses.

Physicists are also trying to determine the nature of dark matter particles by snaring them in detectors on Earth. They've seen nothing so far, but the measurements have tended to rule out extremely heavy particles. "We're approaching the problem from two sides," Belokurov says. "Hopefully one day we'll meet in the middle."

Then again, the detectors might see nothing. The ground-based experiments all make a big assumption: that dark matter interacts with ordinary matter in some as yet unknown way, other than gravity. Many detectors, for example, are designed to pick up a tiny recoil from a dark matter particle's collision with the nucleus of a heavy atom in the detector. But that approach is moot if dark matter interacts
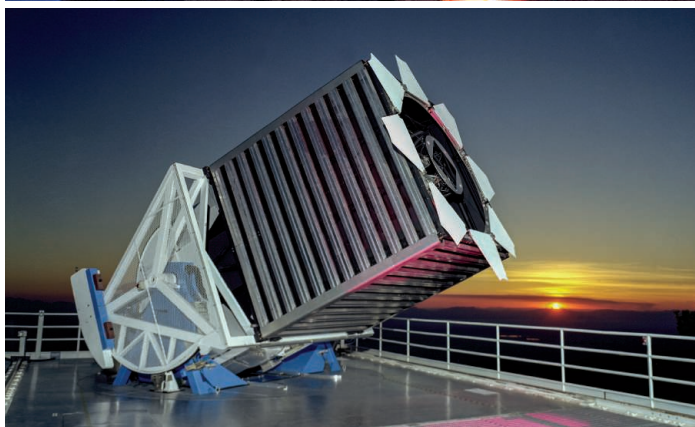
Deep sky surveys have driven a surge in stream discoveries. The Sloan Digital Sky Survey uses a 2.5-meter telescope (bottom left) at Apache Point Observatory in New Mexico. The Dark Energy Survey relies on the 4-meter Blanco telescope in Chile (top). And the 8.4-meter Large Synoptic Survey Telescope is set to open in Chile in 2022 (bottom right).

with ordinary matter only through the force of gravity. Then the only way to narrow down dark matter's properties might be to look for stream gaps caused by subhalos, Erkal says. "That's what I write on all my grant proposals."

THE GAIA SATELLITE will gather data for several more years, pushing out to more distant stars and streams. And new streams are still turning up in the data released on Zero Day. A team led by Rodrigo Ibata, an astronomer at Strasbourg Observatory in France who helped discover Sagittarius, has already found six new streams in the Gaia data. The researchers are working on finding more, using an algorithm that automatically identifies streams. Grillmair knows his days as the stream king are probably numbered. "I will be very shortly eclipsed," he says.

And that's even before the arrival of the Large Synoptic Survey Telescope in Chile, which will turbocharge the search for streams when it begins operations in 2022. With its giant 8.4-meter mirror, it will catalog some 10 billion stars, out to the very edge of the halo. Astronomers are also making headway in planning high-throughput spectroscopy surveys, which would make it possible to assign stars to streams according to their chemical makeup. "It's like DNA matching," Grillmair says.

For their part, Bonaca, Price-Whelan, and Hogg are taking the next step on GD1—doing the detailed modeling to confirm that the gap and spur are the result of a dark matter collision. Bonaca says her latest runs suggest a crash with a subhalo weighing a few million solar masses that occurred some 500 million years ago. But she's still not sure. She wants time on the Hubble Space Telescope to zoom in on the region of the gap and spur, where she hopes to see many faint red dwarfs—common stars, which her current filtering method excludes, that could be members of GD1. By observing across several years, she may be able to obtain proper motions for the red dwarfs, allowing her to decide whether to assign them to the stream.

"We want to go deeper in those regions to be more confident," she says. Ultimately, she might be able to estimate the angle and velocity of the collision. That, in turn, could point to the location of the subhalo today.

As much as it is a golden age for astronomers studying streams, it is also a golden age for the streams themselves. The Milky Way is entering late middle age. Plenty of streams have been scattered into oblivion, but plenty remain in orbit. And hundreds of pristine satellite galaxies await their transformation into streams. The balance is changing, however. Already, dead dwarf galaxies outnumber living ones. Eventually—give it 10 billion years or so—the last streams will dissolve into the halo and lose the memory of their origin. The ghosts will become ghostlier, and the halo will homogenize. The rivers of stars will become one with the sea. ■

INSIGHTS

LETTERS

NEXTGEN VOICES

Quality mentoring

In her Working Life, "Paying it forward as a mentor" (3 August, p. 522), B. Abderrahman describes how a mentor's encouragement can help shape a career. She then explains how her positive mentorship experience inspired her to mentor others. We asked young scientists to **describe one quality of a mentor you've had that you will try to emulate when you become a mentor yourself**. Respondents from around the world wrote in appreciation of their patient, honest, humble, and supportive role models. Excerpts from their responses are printed here. —**Jennifer Sills**

Individualized support

My mentor, like myself, is a first-generation graduate student. She had to pave her way through academia and fight for the projects she knew had value. In a male-dominated field, my mentor had to speak up when she was being overlooked. By creating a path for herself, she taught me how to advocate for science, for others, and for myself.

Lauren Segal
Office of Technology Management,
University of Illinois-Chicago, Chicago, IL 60626,
USA. Email: lsegal3@gmail.com

Good mentors should work to identify their mentees' interests, even when it means stepping out of their own comfort zone. My mentor has spent countless hours guiding me on problems at the intersection of math, philosophy, statistics, and biology, solely because of my interest in them.

Divyansh Agarwal
Department of Statistics, The Wharton School,
University of Pennsylvania, Philadelphia, PA 19104,
USA. Email: divyansh@upenn.edu

When I told my professor that I was planning to commit the cardinal sin (leaving academia for industry), he didn't sour and inform me that scientists can only be successful as professors. Instead, he immediately began searching for industry collaborations and potential internship opportunities, and he suggested that I attend industry-focused conferences. He was mindful of my situation and sought to help, not impede, my career goals.

Kyle J. Isaacson
Department of Bioengineering, University
of Utah, Salt Lake City, UT 84108, USA.
Email: kyle.isaacson@utah.edu

Before I even knew what my interests were, my mentor found the perfect opportunities for me and pushed me to succeed. He suggested paths that I would have never considered myself. When we discussed my hesitation, my mentor assured me that these endeavors would be challenging but rewarding. Having someone identify activities for me outside of my comfort zone and encouraging me to pursue them helped me grow as an individual more than I could have imagined.

Theresa B. Oehmke
Civil and Environmental Engineering, University
of California, Berkeley, Berkeley, CA 94720, USA.
Email: toehmke@berkeley.edu

Academic excellence

My mentors instilled in me discipline in research, strict observance of regulations to prevent lab accidents, and adherence

to project timelines. Their lab meetings helped me assess my progress, solve problems, and properly document results, and regular paper presentations kept me up to date with the latest research and helped me incorporate new ideas.

Brijesh Kumar
Dr. Sneh Lata Singla-Pareek's Lab, Plant Stress Biology Group, International Centre for Genetic Engineering and Biotechnology, New Delhi 110067, India. Email: brijeshkumar2412@outlook.com

My mentor is very supportive, but she is also the best devil's advocate I know. At the start of my Ph.D., I often interpreted results within the framework of my existing hypothesis, sometimes wondering if the experiment had gone wrong if the data did not fit. My mentor would instead challenge my hypothesis, and these discussions have helped me grow into a better scientist.

Jennifer Shuen Chen
Department of Immunobiology, Yale School of Medicine, New Haven, CT 06510, USA. Email: jennifer.s.chen@yale.edu

Humility and empathy

It is empowering to work with a mentor who is open to learning from his or her mentees. I will always strive to be the kind of mentor who never pretends to have all the answers.

Joseph Michael Cusimano
Department of Pharmacy, University of Michigan, Ann Arbor, MI 48109, USA. Email: jjmcu@med.umich.edu

Having the ability to understand other people's concerns without making them feel embarrassed and inadequate is the key to bringing out the best in people. By being empathetic, a mentor gains respect and inculcates a mindful work environment.

Swati Negi
Department of Environmental Systems Science, ETH Zurich, 80092 Zürich, Switzerland. Email: swati.negi@usys.ethz.ch

My mentor led by setting examples of humble excellence, rather than by inducing fear through ultimatums, threats, or castigation. In doing so, she freed me from fears that otherwise would have stifled me.

Irina Tiper
Rockville, MD 20852, USA. Email: irinavtiper@gmail.com

Collaboration and networking

During my training as a biomedical engineer, I was sometimes reluctant to let others work on my project. My professor's open attitude and confidence convinced me to explore collaborations and made me realize that including others in my work not only is personally rewarding but also leads to better and faster scientific accomplishments.

Adrianus J. Bakermans
Department of Radiology and Nuclear Medicine, Amsterdam UMC, University of Amsterdam, 1105 AZ Amsterdam, Netherlands. Email: a.j.bakermans@amc.uva.nl

My mentor added meaning to my work by providing opportunities for me to talk to the physicians and patients who would use the devices we develop. I will emulate this by facilitating interactions between people of diverse perspectives.

Mark Martin Jensen
Department of Bioengineering, University of Utah, Salt Lake City, UT 84112, USA. Email: m.martin.jensen@utah.edu

My supervisor took the initiative to build my network. At conferences, he introduced me to researchers interested in my focus area, and he always invited me to meet local and foreign researchers who visit our university. This taught me that being a mentor means creating opportunities for mentees.

Edmond Sanganyado
Marine Biology Institute, Shantou University, Shantou, Guangdong 515063, China. Email: esang001@ucr.edu

My mentor works to create a diverse team, including lab members of multiple nationalities, races, and genders, and creates equal opportunity by encouraging female lab members to present their work at international conferences. This model has enhanced my social, cultural, and ethical learning.

Syed Shan-e-Ali Zaidi
Gembloux Agro-Bio Tech, Plant Genetics Lab, University of Liège, Gembloux, 5030 Namur, Belgium. Email: shan.e.ali@outlook.com

Impartial advice

When my family or colleagues give me advice, it is biased, unintentionally, by their own interests. In contrast, my mentor helps me to look at problems from a global perspective. He tries to ensure my future success without worrying that a decision may take me far from home or mean leaving my current company.

Carmen Romero-Molina
Department of Biochemistry and Molecular Biology, University of Seville, 41012 Seville, Spain. Email: carmin533@hotmail.com

When I was deciding whether to do a year of volunteer work, I appreciated one professor's clear-cut advice: "Take full advantage of the opportunity you take (volunteering, industry work, or graduate school) regardless of your decision."

Santiago Esteban Martínez
AGP de Colombia, Bogota, Colombia. Email: smarti17@alumni.nd.edu

My greatest mentors have an uncanny ability to hear that far surpasses active listening. Great mentors are rarely those assigned as advisers, committee members, or sanctioned mentors; instead, they are individuals found through networking, shared interests, and happenstance.

Sarah Marie Anderson
Washington, DC 20018, USA. Email: sarah.m.anderson.10@gmail.com

Respect and trust

My scientific mentors were always honest and positive, which helped me to develop resilience. Even in the face of terrible results, they would provide constructive comments. Especially at this moment in Brazil, where the conditions are not favorable for science, I will be forever grateful to them.

Guilherme Martins Santos
Laboratório de Farmacologia Molecular, Department of Pharmacy, University of Brasilia, Brasilia, Distrito Federal, 70910-900, Brazil. Email: gsantos@unb.br

I believe guiding with respect is the most important quality in a mentor. Respect encompasses not only treating kindly, but also truly listening, congratulating, encouraging students to be independent, and trusting our abilities.

Ana Laura De Lella Ezcurra
Instituto de Investigaciones en Medicina Traslacional, CONICET-Universidad Austral, Buenos Aires, B1629WWA, Argentina. Email: anadelella@gmail.com

My mentor gave me unconditional positive regard—she treated me as capable, smart, and likely to succeed, no matter whether things were going well or badly

in the moment. This gave me a safe space to be myself, experiment, and push my limits with less fear of failure.

Janine Farragher
Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB T2N 2T9, Canada. Email: janine.farragher@ucalgary.ca

My mentor has always encouraged me to face my fears. This has helped me not only in the professional sphere but also in my personal life. There was a time when I was ready to give up my Ph.D., but my mentor's unrelenting faith and constant motivation kept me going.

Vandana Sharma
Department of Hematology, All India Institute of Medical Sciences, New Delhi, Delhi 110029, India. Email: dr.vandanasharma.phd@gmail.com

What I valued the most as a trainee was the confidence my mentors had in my ability to make an impact as a scientist. As an underrepresented minority, it was empowering to have mentors who made me feel like I belonged and could hold my own with my peers.

Gregg Duncan
Fischell Department of Bioengineering, University of Maryland, College Park, MD 20742, USA. Email: gaduncan@umd.edu

My mentor has always valued my input and treated me as an equal. This has allowed me to approach problems with a sense of freedom and comfort, giving me confidence to embrace creative solutions without fear of being dismissed.

Ken Dutton-Regester
Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, QLD 4006, Australia. Email: ken.dutton-regester@qimrberghofer.edu.au

Honesty and open communication

My preceptor in my medical course fostered an open environment—a safe place where students could share their thoughts without feeling judged. He was open about personal and uncomfortable topics, such as social anxiety, and offered insight about how to deal with the many challenges I faced.

Sun Ae Kim
University of Central Florida College of Medicine, Orlando, FL 32827, USA. Email: sunkim@knights.ucf.edu

My first mentor shared not only his publication and research goals but also the pressure and difficulties he had to face as an independent researcher. My current mentor shares with me what it is like to lead a group as a woman. Their willingness to be frank and honest with me has been helpful as I navigate my own career.

Sha Yu
School of Biological Sciences, Seoul National University, 08826 Seoul, South Korea. Email: shayu@snu.ac.kr

Patience and freedom to explore

As the boss of a trainee, a mentor needs to provide clear instructions, guidance, and rules. As a senior colleague, the mentor needs to give the junior academic space and opportunities to develop his or her own ideas. A good academic mentor can switch between these two hats and establish a dynamic balance.

Beat A. Schwendimann
Graduate School of Education, University of California, Berkeley, Berkeley, CA 94720, USA. Email: beat.schwendimann@gmail.com

My Ph.D. mentor gave me the opportunity to pursue my own research and is sensitive to the unique needs of training and research in an emerging country. His kindness, generosity, and scientific rigor have shaped my career.

Juergen K. V. Reichardt
Vice Chancellery of Research and Innovation, Yachay Tech University, San Miguel de Urcuquí, Imbabura, 100119, Ecuador. Email: jreichardt@yachaytech.edu.ec

On the day my supervisor assigned my first project to me, he said, "Now you are the owner of this project. I am here to learn from you." Since then, he has been like a curious student, always challenging me to improve. Instead of mentoring me at every step, he allowed me to learn from my own mistakes.

Antarip Halder
Solid State and Structural Chemistry Unit, Indian Institute of Science, Bangalore, Karnataka 560012, India. Email: antarip.halder@gmail.com

My thesis adviser has given me tremendous opportunity to find my own scientific voice, to unravel the narratives that speak to me, and to develop the techniques I feel the field needs. Sometimes I find myself lost in an overly ambitious experiment, with echoes of future I-told-you-sos ringing in my ears, but this creative space is where I learn the most.

Allison F. Dennis
The Program in Cell, Molecular, Developmental Biology, and Biophysics, Johns Hopkins University, Baltimore, MD 21218, USA. Email: adenni16@jhu.edu

Given Brazil's unfavorable science funding landscape, my mentor's creativity was essential for me to carry out my doctoral thesis. He encouraged me to complement my bench work with review articles addressing emerging topics of my research field.

Joel Henrique Ellwanger
Department of Genetics, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, 91501-970, Brazil. Email: joel.ellwanger@gmail.com

When I began my doctoral program, I barely knew a thing about epidemiology. My mentor patiently taught me everything from scratch, from coding, to writing a manuscript, to giving a presentation. He was never pushy and always had his door open when I needed guidance.

Yu-Han Chiu
Departments of Nutrition and Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA 02115, USA. Email: yuc187@mail.harvard.edu

Kangaroos (*Macropus giganteus*) with high-crowned teeth graze in a modern grassland environment.

PERSPECTIVES

EVOLUTION

Marsupial responses to global aridification

Tooth evolution in Australian kangaroos was a late response to climate change in the Neogene

By **P. David Polly**

Perhaps none of Australia's remarkable endemic plants and animals are more iconic than the kangaroos and wallabies that constitute the superfamily Macropodoidea. These pouched animals share a common ancestor with all the world's mammals, but like Australia's other marsupials, they evolved in isolation from small opossum-like founders that inhabited the southern supercontinent of Gondwana in the late Mesozoic [100 to 66 million years (Ma) ago] (*1*). The evolution of these iconic Australian groups (or clades) offers comparative opportunities for understanding how global-scale processes like climate change interact with the exceedingly contingent processes of evolution, adaptation, and extinction. On page 72 of this issue, Couzens and Prideaux (*2*) show that dental evolution in kangaroos responded to global aridification much like it did in other mammalian herbivores around the world, but that

Earth and Atmospheric Sciences, Indiana University, Bloomington, IN 47405, USA. Email: pdpolly@indiana.edu

in Australia, tooth specialization was linked to a late spread of grasslands that postdated the onset of drier habitats.

Couzens and Prideaux's conclusions are based on analysis of the functional traits that link macropodoids to vegetation and thus to climate. Classically, paleontologists have studied biotic responses to global change by measuring taxonomic turnover (*3*). The analysis of taxonomic diversity documents mass extinctions, clade turnovers, and changes in global carrying capacity (*4*), but it is too coarse to study how evolutionary adaptation contributes to these events. Ecometrics, or the analysis of turnover in functional traits, atomizes organisms into form-function pairings between specific traits and associated environmental factors to more directly understand the processes by which taxonomic diversity changes (*5*).

For example, the durability of teeth depends largely on the absolute height of the crown because enamel and dentine have similar properties in all vertebrates. Herbivores that live in environments with more abrasive vegetation should therefore have higher-crowned teeth (hypsodonty) regardless of

whether they are marsupials or placentals (*6*). The tooth crown height (measured as the ratio of crown to root) is thus linked by natural selection to aridity and vegetation cover, especially grasslands and gritty habitats (*7*).

Couzens and Prideaux measured the evolution of tooth crown height in living and fossil macropodoids. To characterize the actual dietary demands on the teeth, the authors measured the wear on the animals' occlusal surfaces (the masticating surfaces of posterior teeth). Using independent proxies for global climate (oxygen isotopes), the spread of grasslands in Australia (pollen and carbon isotopes), and airborne grit (aeolian flux in nearby ocean sediment cores), the authors analyzed the interplay between tooth wear, crown height, and environmental change in the radiation of macropodoids.

Unlike in North America and Eurasia, global aridification was not accompanied in Australia by an increase in herbivore tooth crown height. In North America and Eurasia, hypsodonty (high-crowned teeth) became increasingly common with the onset of drier global climates in the mid Miocene (10 to 15 Ma ago) (*8*, *9*). But in Australia, tooth wear and crown height did not increase until the Pliocene (3 to 4 Ma ago), when grasslands expanded and true kangaroos radiated.

So, what happened 10 million years earlier when Australia became more arid? Instead of browsers (herbivores that fed on leaves, soft shoots, or fruits of plants such as shrubs) giving way to hypsodont grazers (herbivores that fed on grass or other low vegetation), browsing macropodoids got bigger. The sthenurines—giant kangaroo-like animals that weighed as much as 240 kg and strode instead of hopped (*10*)—radiated into the increasingly arid habitats of 14 Ma ago. Their low-crowned teeth, short faces, large body size, and presumably slower mass-specific metabolic rates allowed them to thrive by browsing the increasingly sparse vegetation during the early phase of aridification. Only later, when abrasive, silica-rich grasslands replaced the earlier arid vegetation, did fast-moving true kangaroos (with higher-crowned teeth) replace their slow-moving sthenurine cousins through a combination of clade replacement and trait evolution.

Evolution is a process that is often context specific, rife with contingencies. Stephen Jay Gould wondered what would change if we could "replay life's tape" (*11*). Although comparing evolution on isolated continents does not let us to play the tape again, it does at least allow us to listen in stereo as independent voices respond to the same global theme. Couzens and Prideaux's data dem-

onstrate a disjunction between the onset of aridity and the evolution of high-crowned teeth in Australia that differs from the classic patterns in Eurasia and North America. The Australian transition illuminates a previously identified anomaly in South America in which hypsodonty evolved in clades inhabiting humid forests that predated the spread of grasslands by almost 20 Ma (*12*). High-crowned teeth in South America were favored because the vegetation was gritty from frequent ash falls from the nascent Andes, not because the climate was arid. So, too, in Australia was hypsodonty favored by gritty vegetation, this time grasslands. The classic connection between hypsodonty and climate is at least one step removed from climate (C_4 grasses are favored over many other plant groups in hot, dry climates) and sometimes completely decoupled. Aridity, grasslands, and hypsodonty were entangled in classic North American data, but the comparative histories of mammals in Australia and South America help disentangle the mechanisms of biotic response to environmental change.

Independent lines of evidence for environmental conditions, diet, and tooth morphology are this paper's strength. Functional morphologists studying selection in living organisms routinely control for environmental effects with common garden or other experiments. Paleontologists are often forced to assess cause and effect directly from the morphology of the fossils they study, compromising power to determine whether morphological changes are truly responses to environmental ones. Teeth carry their own independent record of phenotype (the product of genetics and development) and function (impact of environment) because they wear with use rather than remodeling as does almost every other anatomical system. This and other methodological advances like computational finite element analysis now allow paleontologists to conduct persuasive analyses of form and function in the fossil record. ∎

REFERENCES

1. F. S. Szalay, *Evolutionary History of the Marsupials and an Analysis of Osteological Characters* (Cambridge Univ. Press, 1994).
2. A. M. C. Couzens, G. J. Prideaux, *Science* **362**, 72 (2018).
3. J. J. Sepkoski Jr., R. K. Bambach, D. M. Raup, J. W. Valentine, *Nature* **293**, 435 (1981).
4. M. Foote, *Ann. Rev. Ecol. Syst.* **28**, 129 (1997).
5. P. D. Polly et al., *Proc. R. Soc. B* **278**, 1131 (2011).
6. J. Damuth, C. M. Janis, *Biol. Rev.* **86**, 733 (2011).
7. S. J. King et al., *Proc. Natl. Acad. Sci. U.S.A.* **102**, 16579 (2005).
8. C. M. Janis, J. Damuth, J. M. Theodor, *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **177**, 183 (2002).
9. M. Fortelius et al., *Evol. Ecol. Res.* **4**, 1005 (2002).
10. C. M. Janis, K. Buttrill, B. Figueirido, *PLOS ONE* **9**, e109888 (2014).
11. S. J. Gould, *Wonderful Life: The Burgess Shale and the Nature of History* (W. W. Norton, 1989).
12. C. A. E. Strömberg, R. E. Dunn, R. H. Madden, M. J. Kohn, A. A. Carlini, *Nat. Commun.* **4**, 1478 (2013).

10.1126/science.aav1602

SIGNALING

Smoothening out the patches

New roles are discovered for cholesterol transport in a key developmental signaling pathway

By **Anselm Sommer**[1,2] *and* **Mark A. Lemmon**[1,2]

Signaling pathways regulated by the lipidated protein ligand Hedgehog (HH) direct the development of all metazoans through unique molecular mechanisms that remain elusive. More than two decades after their discovery, a series of recently determined structures of two key HH signaling transducers—the transmembrane proteins Patched (PTCH) and Smoothened (SMO)—are now beginning to reveal the signaling events triggered by HH and homologous ligands at the cell membrane. On page 52 of this issue, Qi *et al.* (*1*) describe structures derived by cryo–electron microscopy (cryo-EM) that harmonize recent reports (*2*, *3*) of the structure of the HH receptor PTCH, an

> **"Signaling pathways regulated by…Hedgehog (HH) direct the development of all metazoans…"**

important tumor suppressor (*4*). Together with recent crystal structures of the seven-transmembrane-spanning protein SMO (*5*, *6*), they provide important new insight on molecular events in HH signaling and suggest new opportunities for targeting this pathway in cancer and other diseases (*4*).

SMO was first identified as the transducer of HH signals across the membrane and belongs to the F class of G protein–coupled receptors (GPCRs), along with Frizzleds, receptors in another important developmental signaling cascade—the Wnt pathway. It subsequently became clear, however, that the physical receptor for ligands in the HH family is actually PTCH (*7*) and that these ligands regulate SMO indirectly through PTCH. In the absence of bound HH, PTCH inhibits SMO signaling. This inhibition is relieved upon HH binding to PTCH, in a catalytic manner that does not require direct PTCH-SMO interactions (*8*); this finding suggests

[1]Department of Pharmacology, Yale University School of Medicine, New Haven, CT 06520, USA.
[2]Yale Cancer Biology Institute, West Haven, CT 06516, USA. Email: mark.lemmon@yale.edu

that a mediator is likely to be involved. Because SMO binds cholesterol-like molecules (*5–7*) and PTCH resembles the Niemann-Pick type C1 (NPC1) cholesterol transporter (*8*), sterols have become the prime suspects.

The new cryo-EM structures of human PTCH1 (*1–3*) range in resolution from 3.5 to 3.9 Å and confirm the structural similarity of PTCH1 to NPC1. They detail how the 12 transmembrane α helices (arranged as two groups of six) and the two homologous extracellular domains (ECDs) form a two-fold pseudosymmetric structure. Importantly, cholesterol-like densities are seen in two locations in PTCH1; one is sandwiched between the two ECDs, and the other occupies a membrane-facing cavity in the transmembrane region's sterol-sensing domain (SSD) that is analogous to a cholesterol-sized pocket in the NPC1 cholesterol transporter (*9*). These densities appear to be connected by a contiguous "tunnel" through PTCH1 (and thus through the cell membrane), of sufficient size to accommodate cholesterol (*1*). This tunnel could allow transport of sterol-like molecules from the SSD within the membrane (and/or other intramembrane "gates") to the extracellular space. Indeed, previous studies have implicated PTCH1 directly in cholesterol export (*10*).

The structures also reveal for the first time how the N-terminal signaling domain of the human HH homolog, Sonic HH (SHH-N), binds PTCH1. Initially adding to uncertainties in the field, however, the first two structures (*2*, *3*) showed SHH-N bound to PTCH1 in two quite different ways, using distinct and nonoverlapping parts of the ligand's surface. Unexpectedly, this is resolved in the structure from Qi *et al.* (*1*), which shows a single SHH-N protein bound simultaneously to two PTCH1 molecules.

In PTCH1–SHH-N complexes formed using a "native" (palmitoylated) form of SHH-N (*3*), the palmitoyl chain at the N terminus of SHH-N dominates the interface with PTCH1. The N-terminal ~15 amino acids of the ligand project between the two PTCH1 ECDs. This places the N-terminal palmitoyl moiety in a cavity between the ECDs at the membrane lipid headgroup level, occluding the putative tunnel through the PTCH1 molecule, which could inhibit cholesterol transport. In another structure (*2*), nonpalmitoylated SHH-N binds a similar region on

the PTCH1 surface (mostly involving ECD1) but uses the opposite side of the ligand, including a surface at which divalent cations are bound in all HH structures (11). Gong et al. (2) suggest that this mode of SHH-N binding will also limit sterol egress from the tunnel by "sealing" the ECD sterol binding site. This second mode of PTCH1 binding appears to be the weaker of the two (3), but its importance is demonstrated by the finding that an antibody (5E1) that recognizes precisely the same surface (2, 3) inhibits HH signaling. Conversely, a palmitoylated N-terminal SHH-N fragment of just 22 amino acids can partially activate HH signaling with no contribution from the globular portion of SHH-N (12). The Qi et al. (1) study argues that both interfaces are used, resolving this conundrum. Interestingly, PTCH1 showed a tendency to oligomerize in preparations used for each study (1–3). When Qi et al. (1) added 1 mM Ca²⁺ [not present in their previous study (3)] to occupy the divalent cation binding sites on SHH-N, they observed a complex in which two PTCH1 molecules bind a single native SHH-N, using both of the previously observed interfaces. Mutations affecting only one or the other of the two SHH-N surfaces argue that both interfaces are important for signaling, thereby implicating the 2:1 (PTCH1:SHH-N) complex as a key signaling entity.

> **"The new PTCH1 structures...of this important signaling system...may also suggest therapeutic opportunities."**

The new views of PTCH1 fit nicely with recent advances in understanding SMO regulation and the role of cholesterol as a SMO activator (4–6, 13). A crystal structure of active SMO suggests that it too has a tunnel, open to the membrane at one end and to the extracellular cysteine-rich domain (CRD) at the other (5)—including a region analogous to the agonist-binding site in other GPCRs. This tunnel is large enough to allow cholesterol to transit from the inner leaflet of the membrane to the extracellular agonist-binding site and/or to the CRD, reorienting it and allosterically activating SMO (5, 6).

In the model that seems to be emerging, unliganded PTCH1 locally depletes cholesterol from the inner leaflet of the cell membrane to inhibit SMO signaling activity (see the figure). Indeed, cholesterol is required for SMO activity in a reconstituted system (13), and inhibition of SMO by PTCH1 in this system requires the presence of a transmembrane Na⁺ gradient (13)—consistent with the homology that PTCH1 and NPC1 share with cation gradient–driven efflux pumps and suggesting that cholesterol transport by PTCH1 is chemiosmotically driven (8). When SHH-N binds to PTCH1, the path for cholesterol efflux is thought to be occluded. The resulting inhibition of cholesterol transport by PTCH1 would lead to reaccumulation of cholesterol in the membrane and restoration of SMO signaling to GLI transcription factors.

Numerous important questions remain, and this is only a working model. It remains formally possible that PTCH1 transports an inhibitor that shuts off SMO signaling rather than starving SMO of a key activator by depleting cholesterol. Inhibiting PTCH1-mediated transport would still promote SMO activity in this scenario. Moreover, the importance of having two PTCH1-binding sites on SHH-N is difficult to understand. The resulting bivalence will certainly enhance SHH-N avidity for cell-surface PTCH1. It is not completely clear from all the structures that both interaction modes would inhibit cholesterol efflux equally. Might PTCH1 dimerization also play a signaling role? If inhibition of PTCH1 transporter function is incomplete, the second binding mode could aid signaling by promoting PTCH1 internalization to remove it from the membrane (12). Intriguingly, numerous other HH-interacting proteins that regulate signaling by this ligand bind to the divalent cation-containing surface of HH (11) and may regulate signaling by modulating formation of the 2:1 (PTCH1:SHH-N) complex.

The new PTCH1 structures, beyond enriching our understanding of this important signaling system, may also suggest therapeutic opportunities. Inhibition of deregulated HH signaling with SMO antagonists such as vismodegib is important in treating medulloblastoma and advanced basal cell carcinoma (4). SMO agonists have also been developed, which may have value for regenerative therapies. Mechanistic details of PTCH1 function may suggest additional ways of modulating this pathway. Finally, it is also interesting to speculate that the current model for sterol regulation of SMO, supported by the new PTCH1 structures, might be mirrored in regulation of the closely related Wnt-binding Frizzleds (14) by free fatty acids that bind independently to their CRDs (15). ∎

REFERENCES

1. X. Qi et al., Science **362**, eaas8843 (2018).
2. X. Gong et al., Science **361**, eaas8935 (2018).
3. X. Qi et al., Nature **560**, 128 (2018).
4. F. Wu et al., Cell Chem. Biol. **24**, 252 (2017).
5. P. Huang et al., Cell **174**, 312 (2018).
6. E. F. X. Byrne et al., Curr. Opin. Cell Biol. **51**, 81 (2018).
7. J. M. McCabe, D. J. Leahy, J. Biol. Chem. **290**, 3500 (2015).
8. J. Taipale et al., Nature **418**, 892 (2002).
9. X. Li et al., Proc. Natl. Acad. Sci. U.S.A. **113**, 8212 (2016).
10. M. Bidet et al., PLOS ONE **6**, e23834 (2011).
11. P. A. Beachy et al., Genes Dev. **24**, 2001 (2010).
12. H. Tukachinsky et al., Proc. Natl. Acad. Sci. U.S.A. **113**, E5866 (2016).
13. B. R. Myers et al., Proc. Natl. Acad. Sci. U.S.A. **114**, E11141 (2017).
14. X. Zhang et al., Nat. Commun. **8**, 15383 (2017).
15. P. Chen et al., Science **360**, 664 (2018).

10.1126/science.aav1025

Hedgehog regulation of Smoothened

Insight on the molecular events in HH signaling suggests new opportunities for targeting this pathway in cancer and other diseases.

Without SHH
PTCH1 exports membrane cholesterol through a structurally identified "tunnel" to deplete this lipid, which is required for SMO signaling. SMO is therefore inactive.

Cholesterol export
PTCH1
Membrane
SMO
No signal
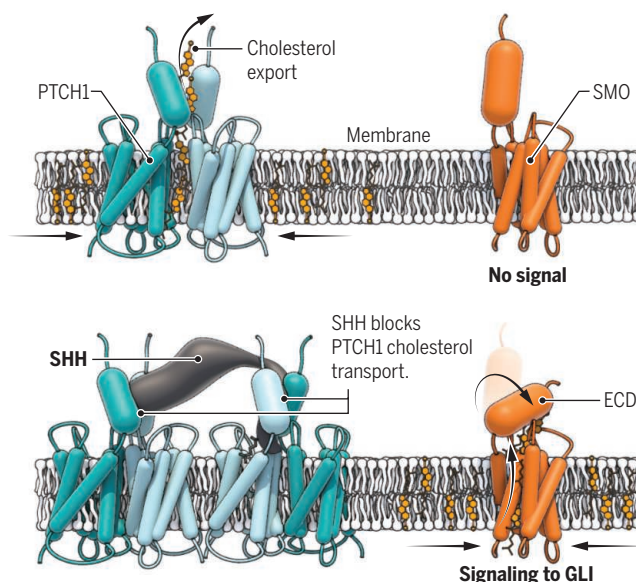
SHH signaling
SHH binds two PTCH1 molecules, occluding their cholesterol "tunnels" in different ways. Export of cholesterol is blocked, allowing it to traverse and activate SMO.

SHH
SHH blocks PTCH1 cholesterol transport.
ECD
Signaling to GLI

GRAPHIC: C. BICKEL/SCIENCE

CHEMISTRY

Activating plasmonic chemistry

Plasmonic photocatalysts can reduce activation barriers and unlock reaction pathways

By **Emiliano Cortés**[1,2]

Industrial applications of photocatalysis have remained elusive, and perhaps the simplest explanation is that the cost-to-benefit ratio of triggering a chemical reaction by using light usually pales in comparison to a purely thermal driven process. For many years, the conversion cycle of sunlight into chemical energy was strictly confined to the involvement of materials such as titanium dioxide and zinc oxide, well-known photocatalysts whose light absorption poorly overlaps the solar spectrum. More recently, however, plas-

monic nanoparticles have been reported that can drive photoreactions and can be engineered to operate efficiently across the ultraviolet, visible, and infrared regions of the electromagnetic spectrum. On page 69 of this issue, Zhou *et al.* (*1*) demonstrate that by illuminating a plasmonic metal nanoparticle, the resulting hot-carriers that are produced can reduce the activation barrier of a chemical reaction by exciting the adsorbed molecules, either electronically or vibrationally. The study proves this for a model system, unlocking the rate limiting step—the kinetic bottleneck—for ammonia decomposition onto a copper-ruthenium alloyed catalyst.

1Blackett Laboratory, Physics Department, Imperial College London, SW7 2AZ, London, UK. 2Nanoinstitute Munich, Ludwig-Maximilians-Universität München, 80539, Munich, Germany. Email: e.cortes@imperial.ac.uk; emiliano.cortes@lmu.de

When light impinges on a plasmonic nanoparticle, it can excite collective oscillations of its conduction electrons, known as surface plasmons. A fraction of the light quanta stored in the plasmons can be re-emitted as light, while another portion can decay into an electron-hole pair. These highly energetic hot-carriers, if efficiently extracted from the plasmonic nanoparticle, can be used to drive novel photochemical reactions.

Modifying the landscape of a chemical reaction with an illuminated plasmonic nanoparticle can trigger reactions that are otherwise thermodynamically and/or kinetically hindered. This enables new degrees of

control in the selectivity and the efficiency of heterogeneous photocatalytic reactions (*1–4*). Zhou *et al.* report a light-dependent activation barrier in their model reaction. They achieve a 75% reduction in the energy needed to convert reactants (ammonia) into products (hydrogen and nitrogen) when exciting at the plasmon resonance of their plasmonic photocatalyst relative to the energy required for the same conversion achieved through external heating to the same surface temperature. This suggests that plasmonic photocatalysis enables a shift to considerably milder reaction conditions while still obtaining the same quantity of products. Equivalently, a light-induced reduction in the activation barrier enables the use of the same reaction conditions between the light and the temperature driven process with far higher efficiencies for the illuminated

case (see the figure). In a simplified picture, part of the energy needed for the reaction to proceed has been provided by the plasmonic photocatalyst, which is capable of efficiently harvesting photons and injecting the energy into adsorbed molecules (the reactants).

Optimization of the energy pathway from photons to molecules requires as a first step the efficient absorption of photons by the photocatalyst. Plasmonic nanoparticles have 10 orders of magnitude higher probability of absorbing a photon than conventional di- and triatomic molecules at visible wavelengths. Subsequent injection of the gathered energy into the molecules is the ultimate goal so as to promote either formation, dissociation, or weakening of molecular bonds at milder reaction conditions (*1–3, 5–7*).

A key role of plasmon resonance excitation, beyond enhanced photon harvesting, has been identified in recent years; the non-radiative decay of the plasmon into a hot electron-hole pair (Landau damping) and the subsequent carrier multiplication through electron-electron scattering—the creation of new carriers by the hot ones colliding in the metal—are pivotal actors in electronically or vibrationally exciting the adsorbed molecule (*1–3, 5, 6*). These processes are known as "desorption induced by electronic transitions" (DIET). Zhou *et al.* postulate that in a plasmonic nanoparticle, the metal-molecule interface (or the nanoparticle-reactant interface) can be treated similarly. Plasmon excitation and decay then provide efficient routes to the harvesting and transfer of energy at metal-molecule interfaces, and they can operate efficiently over the entire solar spectrum.

These previously unknown aspects of plasmon-assisted processes can in turn potentially boost or even modify reaction pathways in solely thermal-driven heterogeneous catalytic reactions that have not yet been tested under the spot of a laser, a light-emitting diode, or concentrated sunlight. However, the most studied plasmonic materials have their *d* band (outermost electrons in the atom) completely filled thus highly suppressing their hybridization with the electronic structure of a wide range of interesting molecules. Nonetheless, the catalytic properties of many partially filled *d*-band metals are well known. An appealing route is the integration of plasmonic cocatalysts (such as gold, silver, copper, and aluminium, among others) into traditional catalysts to promote the absorp-

Plasmonic photocatalysis

Plasmon-derived hot carriers can help in reducing the activation barrier of a chemical reaction due to electronic or vibrational excitation of the adsorbed atoms or molecules.

tion of photons at visible wavelengths and the subsequent excitation of the reactants through the plasmon decay, mediated by the efficient adsorption and desorption of the molecules from the catalyst surface (5).

The interplay between both plasmonic and catalytic properties and the channeling of energy in hybrid photocatalysts has received considerable attention over the past few years (5, 6, 8, 9). These new photocatalysts can induce a synergetic effect in reducing the activation barrier of a chemical reaction—through traditional surface-mediated means and through new plasmon-assisted routes (1, 5, 8). Ultimately, the overall thermodynamic balance of a chemical reaction can be altered, and processes that are endergonic (requiring the absorption of energy) at room temperature and normal pressure might now become spontaneous under illumination (10).

Plasmon-induced electronic or vibrational excitation of reactants, intermediate species, and/or products at visible wavelengths can modify chemical reaction pathways (1, 2). The questions of whether the selectivity of these processes can be externally controlled with the illumination wavelength (11), or the efficiency increased with new interfacial solutions (5, 6, 8) or through reduction in the overall energy requirements (1, 2, 4), are still in their exciting early stages. Further agreement on the extended validity of the proposed mechanisms (1, 6, 7, 11) for different metal-molecule interfaces could assist in the design of better plasmonic photocatalysts. Management of light absorption in traditional catalysts is an open challenge and it can be considered the current rate limiting step for photocatalysis at industrial levels (8). The ability to control both the probability of absorbing photons and the energy of absorbed light by metal plasmonic nanoparticles can open new pathways for the manipulation of chemical reactions (12). Transferring this knowledge into traditional catalytic materials and methods could certainly revolutionize the chemical reaction pathways that we have been able to discover and exploit thus far. ∎

REFERENCES

1. L. Zhou et al., Science **362**, 69 (2018).
2. X. Zhang et al., Nat. Commun. **8**, 14542 (2017).
3. E. Cortés, Adv. Opt. Mat. **5**, 1700191 (2017).
4. Y. Kim, D. Dumett Torres, P. K. Jain, Nano Lett. **16**, 3399 (2016).
5. U. Aslam, S. Chavez, S. Linic, Nat. Nano. **12**, 1000 (2017).
6. K. Wu, J. Chen, J. R. McBride, T. Lian, Science **349**, 632 (2015).
7. E. Kazuma, J. Jung, H. Ueba, M. Trenary, Y. Kim, Science **360**, 521 (2018).
8. S. Chavez, U. Aslam, S. Linic, ACS Energy Lett. **3**, 1590 (2018).
9. D. F. Swearer et al., Proc. Natl. Acad. Sci. U.S.A. **113**, 8916 (2016).
10. Q. Wei, S. Wu, Y. Sun, Adv. Mater., e1802082 (2018).
11. C. Boerigter, R. Campana, M. Morabito, S. Linic, Nat. Commun. **7**, 10545 (2016).
12. S. Simoncelli, Y. Li, E. Cortés, S. A. Maier, Nano Lett. **18**, 3400 (2018).

10.1126/science.aav1133

INFECTIOUS DISEASES

Metropolitan versus small-town influenza

Analysis of infectious disease data reveals the driving factors of infection dynamics

By **Jacco Wallinga**[1,2]

Every year, influenza epidemics occur during winter in regions with temperate climates. The duration, size, and precise timing of influenza epidemics vary from season to season—the numbers of hospitalizations and deaths can be substantial. The influenza epidemic in the winter of 2017–2018 in the United States was extraordinarily severe and long, straining the health care system capacity. Various external factors have been proposed to explain the duration, size, and precise timing of seasonal influenza epidemics, including weakened immune systems during winter, increased social crowding indoors, school leave during the Christmas holiday, and the lower humidity in winter. Inferring the relative importance of each of those factors is difficult because they act on a dynamic process that responds sensitively to changes and that is only observable with a measure of error (1–4). How can we determine what is driving influenza epidemics? On page 75 of this issue, Dalziel et al. (5) demonstrate how influenza epidemics in a city are driven by fluctuations in humidity, modulated by the population size. This im-

plies that influenza control measures could work differently in large metropolitan areas compared to small towns.

The unexpected role of city population size in driving influenza epidemics has been revealed through analyzing an impressive, richly detailed dataset that contains the weekly numbers of U.S. patients presenting at doctors' offices with influenza-like illness. Such data are subject to observational biases and underreporting because, for example, not all of these patients with influenza-like illness have been infected with the influenza virus (6). The patient visits in this dataset were recorded over six winters in about 600 U.S. three-digit zip code regions; a three-digit zip code region roughly corresponds to a city. The large number of cities and seasons makes the dataset well suited to examine infectious disease dynamics across a wide range of climate conditions, city sizes, levels of crowding, household incomes, and vaccination coverages. Each of these factors could modify the infectivity or degree of susceptibility of individuals. Subtle fluctuations over time of a driving factor often result in substantial variations in the number of influenza cases, yet the latter are often out of step with the former. Therefore, naïve statistical testing for an association between a factor and the number of observed cases would yield little.

The study of Dalziel et al. avoids the problems of such naïve statistical testing by first reducing the data series of weekly

[1]Center for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, Netherlands. [2]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands. Email: jacco.wallinga@rivm.nl

influenza-like illness cases for each city to a single summary statistic. This summary statistic is then compared against the candidate factors: humidity, population size, residential crowding, median household income, and vaccination coverage. This method only works when the summary statistic captures the relevant features in the data for determining the dynamics. Of the wide range of statistics available (7), the summary statistic of choice is "epidemic intensity." Although the term "intensity" is frequently used to refer to severity or the number of cases in an influenza epidemic, the meaning here is different: A low epidemic intensity reflects a diffusely spread epidemic during the season, whereas a high epidemic intensity reflects a sharply peaked influenza epidemic. The data reveal that the epidemic intensity does not vary with the mean proportion vaccinated at state level. As expected, epidemic intensity is high in cities with larger fluctuations in humidity

> "...metropolitan areas should focus on reducing influenza spread,...small towns should focus on reducing harm."

(8). Importantly, Dalziel et al. find that the epidemic intensity is higher in cities with smaller population sizes, lower levels of residential crowding, and lower levels of median household income.

A potential explanation is that at smaller population sizes, the role of humidity becomes more relevant. Further statistical analysis of the observations offers ample support for such an explanation. This analysis requires a careful approach to statistical modeling, again to avoid the pitfall of naïve statistical testing (9–11). The explanation implies that metropolises (densely populated urban areas with high average household incomes) act as influenza transmission hubs: More people meet and the transmission rates of influenza are higher, leading to influenza epidemics that are more spread out over time. This also implies that in small towns (areas with low population sizes and low average household incomes), the transmission rates of influenza are lower and the influenza epidemics are more sharply peaked. This is important for policy-makers because it indicates that metropolitan areas should focus on reducing influenza spread, whereas small towns should focus on reducing harm.

As the field of infectious disease dynamics becomes increasingly data-driven, the epidemic intensity–based approach has considerable prospects for future research. Perhaps most pressing is the assessment of the effectiveness of vaccination programs and other control measures (12). The proposed approach might also help us to understand the seasonality of epidemics for other infections (13). Focusing on influenza specifically, the outcome of this large natural experiment should be confirmed for other populations, using other datasets, and at different scales. This would ideally include other data, such as age at infection, and other data types, such as serological data, influenza virus subtypes, or even viral genome sequences. Further work needs to be done to test whether the factors are truly the underlying causes; the findings presented here do not rule out that other driving factors of influenza epidemics exist. Similarly, further work is needed to test whether the factors are truly helpful in improving predictions (14) and in improving projections of the impact of alternative vaccination strategies.

The study of Dalziel et al. implies a shift in thinking about infection control: Rather than target public health campaigns to the population at large, there could be benefit in a refined approach that identifies and targets specific small towns or metropolitan areas for control. Summary statistics, such as epidemic intensity, help us to identify which places require more surge capacity to deal with the peak health care demand. They also help us to find good locations for active influenza surveillance where long transmission chains of influenza occur, and where new genetic variants of the influenza virus can be detected. Understanding how dynamics differ between cities enables the development of more accurate short-term, small-scale forecasts of the expected health care demand in a season. Most important, it allows long-term projections that reveal how the shifting demography, growth of cities, and the changing climate alter the infection dynamics and required control efforts. ∎

REFERENCES

1. S. Cauchemez et al., Nature 452, 750 (2008).
2. E. L. Ionides et al., Proc. Natl. Acad. Sci. U.S.A. 103, 18438 (2006).
3. S. N. Wood, Nature 466, 1102 (2010).
4. C. J. E. Metcalf et al., Proc. R. Soc. B 284, 20170901 (2017).
5. B. D. Dalziel et al., Science 362, 75 (2018).
6. W. Yang et al., Proc. Natl. Acad. Sci. U.S.A. 112, 2723 (2015).
7. T. S. Brett et al., J. R. Soc. Interface 14, 20170115 (2017).
8. J. Shaman, M. Kohn, Proc. Natl. Acad. Sci. U.S.A. 106, 3243 (2009).
9. Y. Xia et al., Appl. Statist. 54, 659 (2005).
10. M. Farah et al., J. Am. Stat. Assoc. 109, 1398 (2014).
11. D. E. te Beest et al., Am. J. Epidemiol. 178, 1469 (2013).
12. M. van Wijhe et al., Lancet Infect. Dis. 16, 592 (2016).
13. V. E. Pitzer et al., PLOS Pathog. 11, e1004591 (2015).
14. L. Held et al., Statist. Med. 36, 3443 (2017).

10.1126/science.aav1003

CANCER

Cancer origins—genetics rules the day

Similar cancers from different source tissues share molecular mechanisms

By **Michael S. Kareta**[1,2,3] and **Julien Sage**[4]

A major goal of cancer research is to identify central molecular and cellular mechanisms underlying the development of tumors and their response to treatment, with the aim of uncovering key vulnerabilities. Early events in the development of cancer may inform such vulnerabilities (1), but early tumors are much more difficult to observe and study in patients than established tumors. Indeed, one of the hardest issues to resolve in early tumor development is the relative contributions of the oncogenic driver mutations and the nonpathogenic gene networks expressed in a precancerous cell. On page 91 of this issue, Park et al. (2) investigate the mechanisms of development of neuroendocrine cancer in the lung and the prostate using human epithelial cells in culture. They find that these neuroendocrine tumors can arise from non-neuroendocrine epithelial cells, which converge upon reprogramming toward a neuroendocrine fate via a common and specific combination of genetic factors.

Human neoplasms with neuroendocrine features can originate in nearly all human organs and tissues (3). Often, neuroendocrine cancer cells are classified as "small" cells owing to their morphology. Among tumors with neuroendocrine features, small cell lung carcinoma (SCLC) is a common (~15% of all lung cancer cases) and extremely lethal form of lung cancer (4). SCLC is thought to arise from a rare population of neuroendocrine lung epithelial cells, but other cell types–of-origin have not been excluded (5). SCLC can

[1]Genetics and Genomics Group, Sanford Research, Sioux Falls, SD, USA. [2]Department of Pediatrics, University of South Dakota Sanford School of Medicine, Sioux Falls, SD, USA. [3]Department of Chemistry and Biochemistry, South Dakota State University, Brookings, SD, USA. [4]Departments of Pediatrics and Genetics, Stanford University, Stanford, CA, USA. Email: julien.sage@stanford.edu; michael.kareta@sanfordhealth.org

also arise from lung adenocarcinomas that gain neuroendocrine features as they evolve resistance to treatment (*6*). Small cell prostate carcinoma (SCPC) is a rare form of neuroendocrine prostate cancer at the time of initial diagnosis (~2% of all prostate cancer cases), but this cancer type is increasingly being considered to originate during the progression of prostate adenocarcinoma that becomes resistant to androgen-deprivation treatment (*7, 8*). Similar to SCLC, SCPC often becomes highly metastatic and chemoresistant and has poor patient survival rates (*9*). Both of these tumors are currently treated with platinum-based chemotherapies and radiotherapy; however, because of the rapid acquisition of resistance, alternative treatments are needed. One avenue toward discovering targeted therapies for these tumors is to elucidate their genetic underpinnings. Both SCLC and SCPC typically display loss of function of two essential tumor suppressor genes, retinoblastoma 1 (*RB1*) and tumor protein 53 (*TP53*, which encodes p53) (*10, 11*). However, a broader understanding of the genetic drivers of these tumors and the contribution of host-cell genomics are needed to advance the identification of new targeted therapies.

Park *et al.* find a common set of oncogenic factors that reprogram non-neuroendocrine human prostate and lung epithelial cells into SCPC and SCLC, respectively. They had reported that overexpression of the oncoproteins N-MYC and myristoylated AKT1 (myr-AKT1) in normal human prostate epithelial cells could partially transform these cells into SCPC (*12*). On the basis of frequent alterations in human SCPC, they now report that transduction of normal human prostate epithelial basal cells to express dominant-negative p53, myrAKT1, c-MYC, and the apoptosis inhibitor BCL2 and to inhibit the expression of RB1 (together called PARCB) leads to the development of SCPC tumors upon injection into immunocompromised mice (see the figure). In this system, overexpression of c-MYC and myrAKT1 is indispensable for tumor formation, whereas down-regulation of p53 activity and RB1 is essential for the acquisition of neuroendocrine features.

As SCPC and SCLC are defined by similar histological characteristics and molecular changes, Park *et al.* also analyzed SCLC. They found that SCLC samples were similar to SCPC samples and PARCB-induced tumors, suggesting that, overall, they share a similar transcriptome. Indeed, they were able to transform normal human bronchial epithelial cells into SCLC using the PARCB factors. Furthermore, the PARCB cells and SCLC and SCPC cell lines were more similar to each other than to the epithelial cells they were derived from, demonstrating convergence of pathways leading to neuroen-

A common pathway to small cell neuroendocrine cancer

Park *et al.* identified a selection of oncogenic drivers (PARCB) that can reprogram human non-neuroendocrine prostate and lung epithelial basal cells into neuroendocrine prostate and lung tumors that are similar to SCPC and SCLC.
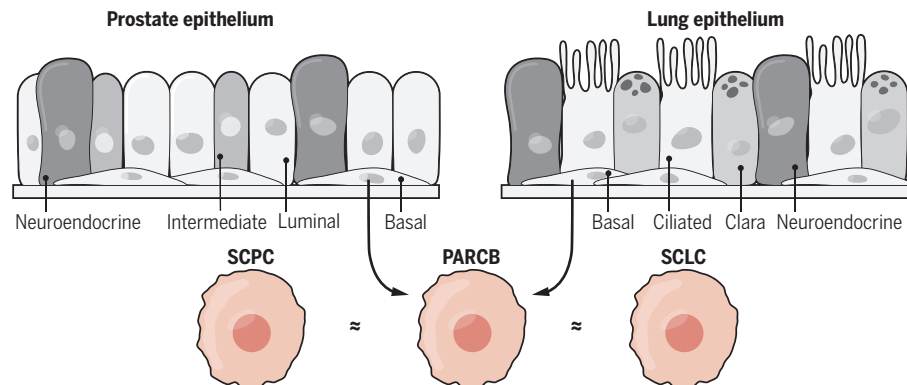
docrine cancer. They identified additional candidate factors that may govern these pathways, including the well-known neuroendocrine fate-determining basic helix-loop-helix (bHLH) transcription factor achaete-scute homolog 1 (ASCL1) (*13*).

This reprogramming methodology to investigate the basic tumorigenic mechanisms of human small cell neuroendocrine cancers poses some interesting questions. As the lung and prostate non-neuroendocrine epithelial cells arrive at a similar neuroendocrine cancer phenotype, it would be interesting to investigate if they follow a common pathway during reprogramming and transformation. One could envision that this reprogramming first imparts cancerous development followed by a gain of common neuroendocrine identity. Alternatively, the epithelial cells from different sources may find distinct paths to a common cancer state. Gene expression analysis at different time points during transformation, possibly at the single-cell level, may help address this point in the future.

Furthermore, does the constant expression of the PARCB factors restrict the plasticity of the resulting cancer cells, or is this oncogenic combination directly responsible for the remarkable ability of small cell neuroendocrine tumors to evade the effects of the many therapies that have been tested in the clinic? If a wide range of neuroendocrine tumors share common mechanisms of development, this would indicate that neuroendocrine tumors could be treated similarly in the clinic, independent of their tissue of origin (*3*). An interesting experiment would also be to use the same PARCB approach with other epithelial cells from other organs and tissues. Not all neuroendocrine tumors express ASCL1, but it is possible that these tumors share some mechanisms with other bHLH factors that may reprogram cells towards a neuroendocrine phenotype. Finally, we now

know that neuroendocrine tumors can arise from three sources, normal neuroendocrine cells, normal non-neuroendocrine cells, and non-neuroendocrine tumor cells. A complete characterization of tumors derived from these three sources would further define the genetic similarities and potential differences within the neuroendocrine tumor spectrum, which is important because there are increasing numbers of patients with epithelial tumors that become resistant to treatment and emerge as small cell neuroendocrine cancer.

Work in mouse models indicates that the same lung cancer subtype may arise from different epithelial cell types–of-origin (*14, 15*). Park *et al.* have now shown that a set of defined oncogenic factors is able to transform different human lung and prostate non-neuroendocrine epithelial cells into neuroendocrine small cell cancers that resemble clinical samples. This ex vivo system is a powerful new tool to model and study cancer development in human cells. ∎

REFERENCES AND NOTES

1. J. E. Visvader, *Nature* **469**, 314 (2011).
2. J. W. Park *et al.*, *Science* **362**, 91 (2018).
3. G. Rindi *et al.*, *Mod. Pathol.* 10.1038/s41379-018-0110-y (2018).
4. J. K. Sabari *et al.*, *Nat. Rev. Clin. Oncol.* **14**, 549 (2017).
5. D. Yang *et al.*, *Cancer Discov.* 10.1158/2159-8290.CD-17-0987 (2018).
6. L. V. Sequist *et al.*, *Sci. Transl. Med.* **3**, 75ra26 (2011).
7. P. Mu *et al.*, *Science* **355**, 84 (2017).
8. S. Y. Ku *et al.*, *Science* **355**, 78 (2017).
9. R. Nadal *et al.*, *Nat. Rev. Urol.* **11**, 213 (2014).
10. J. George *et al.*, *Nature* **524**, 47 (2015).
11. H. Beltran *et al.*, *Nat. Med.* **22**, 298 (2016).
12. T. Stoyanova *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 20111 (2013).
13. D. W. Ball, *Cancer Lett.* **204**, 159 (2004).
14. K. D. Sutherland *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 4952 (2014).
15. G. Ferone *et al.*, *Cancer Cell* **30**, 519 (2016).

SOLID-STATE PHYSICS

Quantum oscillations in an insulator

Even without a Fermi surface, a Kondo insulator exhibits magnetoresistance oscillations

By **N. P. Ong**

A major triumph of the quantum theory of solids was the explanation (Bloch-Wilson theory) of the vast difference in electrical resistivity (14 orders of magnitude) between insulators and metals. In a metal, the conduction electrons define a Fermi surface, whose existence leads to quantum oscillations in resistivity versus a magnetic field. Insulators, with bands completely filled, have no Fermi surface, so quantum oscillations should be absent. Recently, this deeply rooted distinction has been challenged by experiments (*1*). However, the interpretations were muddled by the existence of metallic states on the surface (*2*) and other factors. On page 65 of this issue, Xiang *et al.* (*3*) report torque and magnetoresistance experiments on the "Kondo insulator" YbB$_{12}$, which becomes a bulk insulator below 10 K. They present a persuasive case for a bulk origin for the quantum oscillations detected in the resistance.

In Bloch's theory, the quantum states in all crystalline solids are grouped into bands separated by an energy gap (the states are represented as pixels **k** in momentum or **k** space). In an insulator, the electrons fill the highest occupied band completely. A large energy gap precludes transitions of electrons to the higher-energy conduction band. Hence, an insulator is unresponsive to weak electrical probes. By contrast, a metal has enough electrons to "spill" into the conduction band, which becomes partially filled. A feeble electric field suffices to induce spreading of the occupied states to unoccupied ones, which translates to the flow of a charge current in the metal. Each time we flip on a light switch, we are witnessing this inherently quantum-mechanical process in action.

At 0 K, the occupied states in **k** space in the conduction band collectively define a surface called the

Department of Physics, Princeton University, Princeton, NJ 08544, USA. Email: npo@princeton.edu

Fermi surface. All **k** states outside the Fermi surface are unoccupied. In a magnetic field, electrons execute closed cyclotron orbits in real (physical) space. If the magnetic field is strong, its quantum nature becomes manifest: The field is actually composed of a dense array of a fundamental quantity called the flux quantum (see the figure, arrows in top panel). In a metal, this quantum granularity has a profound effect on the electrons. In real space, a cyclotron orbit (see the figure) is allowed only if it encloses exactly *n* flux quanta, where *n* is an integer. Equivalently, the wave function (see the figure, thick curve

Magnetically driven electrons

Conduction electrons in a solid are driven in cyclotron orbits by applied magnetic fields. The quantum states that result are illustrated.

Real-space orbits
An allowed cyclotron orbit enclosing six flux quanta (arrows). Equivalently, the wave function (thick curve) winds six times to complete the orbit.

Flux quantum

Wave function
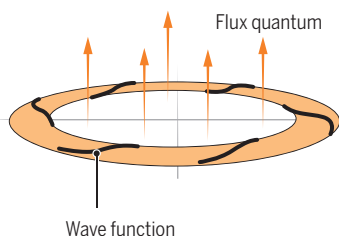
Momentum-space levels
Nested cylinders representing allowed states in momentum (**k**) space. The outermost cylinder is on the verge of emptying.

Landau levels

Fermi surface in zero field

Worm-hole Fermi surface
The angular variations of the resistivity (Shubnikov–de Haas oscillations) measured by Xiang *et al.* in the insulator YbB$_{12}$ imply a Fermi surface of this shape.

in top panel) winds exactly *n* times around the closed loop to return to its starting value (otherwise it is multivalued). All orbits that share the same integer *n* may be grouped together. In **k** space, the groupings correspond to a series of nested cylinders, with their common axis parallel to the field direction (see the figure, middle). The collection of all states in the *n*th cylinder defines the *n*th Landau level. Although the Fermi surface has faded away, its residual imprint serves to determine the occupancy of the Landau levels. A Landau level is partially occupied if its cylinder intersects the imprint volume. Otherwise, it is completely unoccupied.

As the magnetic field increases, the cylinders expand. When a cylinder expands beyond the imprint volume, the abrupt emptying of that Landau level leads to a sharp dip in the observed resistance *R* and the magnetization *M*. Hence, in an increasing field, the sequential emptying of successive levels produces an oscillatory field profile in both *R* and *M*. Plotted versus 1/*B*, the peaks are spaced periodically with a period dictated by the Fermi surface caliper area. The technique is especially valuable when the Fermi surface is elongated or has multiple pockets. By tilting the field direction, the caliper area can be tracked with the changes with the tilt angle of the field.

These quantum oscillations should not be present in an insulator. The material YbB$_{12}$ investigated by Xiang *et al.* is a mixed-valence Kondo insulator. The 4*f* electrons form a very narrow band that intersects a broad band occupied by the 5*d* electrons. Near room temperature, the 4*f* electrons are strongly localized on the Yb ions to form fluctuating local moments. Below 10 K, hybridization between the two bands creates a small energy gap (~15 meV) (*4*). Because the chemical potential falls within this small gap, a nonmagnetic Kondo insulating ground state is realized.

Combining torque magnetometry with measurements of the resistivity, Xiang *et al.* initially detected the onset of quantum oscillations in YbB$_{12}$ near 37 T. They subsequently tracked

oscillations in both R and M to 60 T using pulsed magnetic fields. Their finding appears to be the first report of quantum oscillations in the resistivity ("Shubnikov–de Haas," or SdH oscillations) in an insulator. Repeating the SdH measurements with the magnetic field tilted at an angle, Xiang *et al.* found that the angular variation of the SdH period is distinctly three-dimensional, which they argue excludes a surface origin for the SdH oscillations. The angular variation appears to describe a Fermi surface that resembles a worm hole (see the figure, bottom panel).

Theories for quantum oscillations in an insulator (5–7) may now be confronted in part by experiment. As the temperature is raised, increased scattering of the carriers by lattice vibrations leads to broadening of the Landau levels, which results in an exponential decrease in the oscillation amplitude. Typically, SdH experiments find close agreement with the Lifshitz-Kosevich (LK)

> *"Xiang et al. found that the angular variation of the SdH period is distinctly three-dimensional, which they argue excludes a surface origin for the SdH oscillations."*

expression, which enables the effective mass of carriers to be extracted. Likewise, Xiang *et al.* found good agreement between the damping of the SdH signal with the LK expression, which is in contradiction with some of the predictions.

Last, the quantum oscillations observed with magnetometry ("de Haas–van Alphen" oscillations) display a different behavior in the tilt-angle experiment as well as in the damping rate of the amplitude (8). Xiang *et al.* associate different origins for the two sets of oscillations. Although the new results exclude some classes of proposed theories, many aspects of the experiment pose a stiff challenge to conventional wisdom. ∎

REFERENCES

1. B. S. Tan *et al.*, *Science* **349**, 287 (2015).
2. G. Li *et al.*, *Science* **346**, 1208 (2014).
3. Z. Xiang *et al.*, *Science* **362**, 65 (2018).
4. M. Okawa *et al.*, *Phys. Rev. B* **92**, 161108(R) (2015).
5. M. Dzero, K. Sun, V. Galitski, P. Coleman, *Phys. Rev. Lett.* **104**, 106408 (2010).
6. J. Knolle, N. R. Cooper, *Phys. Rev. Lett.* **118**, 096604 (2017).
7. D. Chowdhury, I. Sodemann, T. Senthil, *Nat. Commun.* **9**, 1766 (2018).
8. H. Liu *et al.*, *J. Phys. Condens. Matter* **30**, 16LT01 (2018).

10.1126/science.aau3840

Preserving microbial diversity

Microbiota from humans of all cultures are needed to ensure the health of future generations

By **Maria G. Dominguez Bello[1], Rob Knight[2], Jack A. Gilbert[3], Martin J. Blaser[4]**

Since World War II, there have been dramatic increases in metabolic, immune, and cognitive diseases, including obesity, diabetes, asthma, allergies, inflammatory bowel disease, and autism. Their incidence has risen, first in the industrialized world and more recently in developing countries (1). In addition to the health effects, there are enormous costs of these diseases: Obesity costs $2.0 trillion and diabetes costs $1.3 trillion per year globally (1–3). As these diseases advance in developing countries, the problem is worsening rapidly. The cost, to health and economies, is becoming unsustainable, with care of chronically ill adults competing with the proper care for the next generation. Are all these distinct diseases independent, or is there a common underlying factor? We believe that changes in the human microbiota occurring concomitantly with industrialization may be the underlying factor. The changes involve the loss of our ancestral microbial heritage to which we were exposed through millions of years of evolution.

The microbiota (and its collective genomes, the microbiome) encompasses bacteria, archaea, viruses, fungi, and other microeukaryotic colonizers that live in or on our bodies. They influence essential host processes, including nutrition, immunity, hormone activity, gut permeability, and neurochemistry. The microbiome, our "other genome," is largely passed from generation to generation, in early life, from mothers to their children (4). The germline and somatic genomes, microbiome, and external environment are all essential features of population variation that are useful for predicting host disease and health outcomes (5, 6). Microbes acquired in early life play crucial roles in guiding the development of immune, metabolic, and neural systems in animal models (7–10) and may also be important in human development, as indicated by epidemiological evidence. Animal studies have shown causation in the relationship between associations of microbiome perturbations in early life and later development of diseases such as obesity, juvenile diabetes, and asthma (7–10).

Industrialization is substantially correlated with reduced human microbiota diversity (11). The gut diversity of South American Amerindians is ~2-fold that of healthy people in the United States (12, 13). Comparing disparate traditional societies with industrialized peoples indicates that the loss of gut microbial diversity is associated with industrialization, not with particular diets, ethnicity, or geography. People in industrialized and developing countries from across the world can be classified into two overall groups: low and high intestinal bacterial diversity, respectively (11–13). Industrialization encompasses many influences on the microbiota, including a highly processed water supply, refined diets, and altered environmental exposure—which may affect the microbiota through effects on the immune system—and the presence of medical care, including pre-, peri-, and postnatal antibiotics; cesarean section delivery; and bottle-feeding—all of which reduce the transmitted and maintained microbial diversity, especially during the critical window of early-life development (see the figure). Loss of microbiota diversity opens up niches for opportunistic invaders, which often do not have the same coevolved constraints. Intestinal bacterial taxa of *Desulfovibrio*, *Bacteroides*, *Prevotella*, *Lactobacillus*, *Oxalobacter*, and lineages in the families Succinivibrionaceae, Paraprevotellaceae, and Spirochaetaceae (11, 12) are disappearing with industrialization. For example, *Oxalobacter formigenes* uses indigestible oxalate and its loss (12) may be hindering protection against the formation of kidney stones, which are composed of calcium oxalate. This is just the beginning of our knowledge about the impacts of living in an industrialized world— we need to better understand which strains in human populations are diminishing and what the functional and pathological implications are for these losses.

In many ways, the loss of our microbial diversity resembles climate change: Industrialization has led to substantial unintended pressures on global ecology. Currently, over 50% of the world's population lives in urban environments, and cities are likely to ab-

[1]Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ 08901, USA. [2]Center for Microbiome Innovation, University of California San Diego, CA 92093, USA. [3]Department of Surgery, University of Chicago, Chicago, IL 60637, USA. [4]Departments of Medicine and Microbiology, New York University Langone Medical Center, New York, NY 10016, USA. Email: mg.dominguez-bello@rutgers.edu

sorb future population growth. We suspect that the microbes disappearing in urban societies are those that are needed to maintain health and prevent many metabolic, immune, and cognitive diseases. If microbial disruption due to urbanization increases the diseases of industrial societies, then the current global pandemics will worsen, with economic impact jeopardizing health care systems.

However, much can be accomplished through better stewardship: curtailing the overuse of antibiotics, limiting cesarean sections to those that are necessary and not for convenience, promoting breastfeeding, removing antibacterial compounds from everyday activities, and changing diets to emphasize the nutrients and foods that promote microbial diversity and metabolism that benefits our health (*14*). But, these efforts will likely only diminish the rate of decline. We must examine ways to accelerate microbial restoration to recover functional activities, those reduced through microbiota depletion or those that, like weeds, overgrew in an affected microbiota, suppressing normal activities and health. We can replace specific microbial gut functions through the targeted provision of microbial metabolites (for example, intestinal butyrate to signal to and nourish colonocytes, or vaginal lactate to select for acid-tolerant microbes) to replace the fermentative activities of endogenous microbiota. Further scientific exploration will enable the development of dietary additives to help replace our lost chemistry. But nature is hard to micromanage because individual organisms have complex functions, with (often overlooked) context-dependent characteristics. This might explain the utility of fecal microbiota transplantation (FMT), a restorative technique for returning ecological balance to the colon, including eliminating life-threatening *Clostridium difficile* infections (*15*), and is being tested for other conditions associated with altered microbiotas. Identifying the activities of specific microbial species and strains is essential for more targeted and effective interventions, such as the provision of specific bacteria or probiotics.

Controlled restoration of the missing microbiota must be balanced against the importance of not introducing harmful microbes. Additionally, there needs to be greater knowledge of the specificity of microbe-host interactions; for example, does

Declining human intestinal microbial diversity with industrialization

There has been a progressive decline in human gut microbiota diversity with industrialization. The compounded effects of chlorinated water, antibiotics, antiseptics, cesarean section birthing, and formula feeding may all contribute. This decline has been linked to the rise of modern diseases: obesity, asthma, food allergies, diabetes, inflammatory bowel disease, and cognitive disorders.
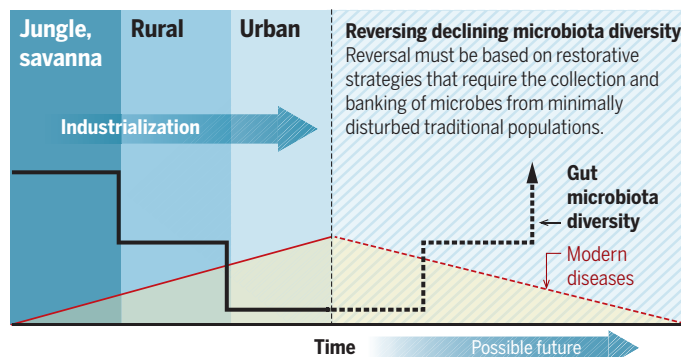
a proposed intervention yield the same outcome in Dublin and Dhaka? Identifying nutrients that specifically enhance the colonization of the selected microbes to be restored, yet maintaining the desired diversity, will be challenging. Moreover, restoration may be dependent on host age, gender, and/or genotype, meaning that finding universal treatments that are broadly effective may be difficult. Fiber (*14*) and other nutrients present in the diet of peoples with more traditional lifestyles may represent some of the key compounds to sustain the microbes that are needed for restoration. The antigenic and metabolic properties of microbes may have been selected based on specific host phenotypes or environments. If a persons' ancestors were milk-drinking cattle farmers, then their microbial exposures, immune responses, and endogenous microbe-immune interactions might differ radically from people whose ancestors were fishermen. Reducing microbial extinctions due to industrialization is important but unlikely to be sufficient to arrest the current disease trend without restoration. Purposeful microbial restoration emphasizes prevention rather than treatment and is possibly the only way that microbial losses can be reversed.

Most urgently, we need to preserve the diversity of ancestral microbes from globally diverse human populations and especially include those who have had the least exposure to urbanization. Using current technology, and under the precautionary principle (to avoid the introduction of products and processes the ultimate effects of which are unknown), it is paramount that we expand the efforts to capture and preserve the human microbiota while it still exists. This is a needed step toward restoration and could help mitigate the potential risk to hu-

man health that urbanization encompasses.

Biobanking initiatives are arising in universities and companies, but most are focused on preserving specimens and strains from individuals in industrialized societies. These collections are prepared for custom microbiome analysis and for clinical uses, such as FMT. Research collections that contain microbes from geographically diverse individuals have begun, but few have collected high-diversity microbiotas from traditional peoples in developing countries, which hold promise for disease-prevention restoration strategies. Such biobanking has challenges, including accessing remote peoples, relying on local teams of scientists who hold the required permits, and training local microbiologists to foster collections, as well as our limited capacity to culture a wide diversity of microbiota species and to classify it. Biobanking efforts need to be supported and expanded into developing countries (that hold the richest human microbial biodiversity) (*11–13*). This will involve sharing and transferring the required technologies from industrialized countries (that hold depleted human microbiotas) to the ancestral microbe-rich developing countries. A global repository of human-associated microbes should back up existing research collections, similar in principle to the inspiring example of the Seed Vault established in the permafrost of Svalbard Island in Norway to preserve the natural biodiversity of plants. We owe future generations the microbes that colonized our ancestors for at least 200,000 years of human evolution. We must begin before it is too late. ∎

REFERENCES AND NOTES

1. C. Bommer *et al.*, *Lancet Diabetes Endocrinol.* **5**, 423 (2017).
2. R. Dobbs *et al.*, "Overcoming obesity: An initial economic analysis" (McKinsey Global Institute, 2014).
3. D. E. Bloom *et al.*, "The global economic burden of non-communicable diseases" (World Economic Forum and the Harvard School of Public Health, 2011).
4. M. Yassour *et al.*, *Cell Host Microbe* **24**, 146 (2018).
5. A. Spor *et al.*, *Nat. Rev. Microbiol.* **9**, 279 (2011).
6. D. Rothschild *et al.*, *Nature* **555**, 210 (2018).
7. K. A. Martinez 2nd *et al.*, *Sci. Adv.* **3**, eaao1874 (2017).
8. A. E. Livanos *et al.*, *Nat. Microbiol.* **1**, 16140 (2016).
9. L. M. Cox *et al.*, *Cell* **158**, 705 (2014).
10. M. C. Arrieta *et al.*, *Sci. Transl. Med.* **7**, 307ra152 (2015).
11. S. A. Smits *et al.*, *Science* **357**, 802 (2017).
12. J. C. Clemente *et al.*, *Sci. Adv.* **1**, e1500183 (2015).
13. T. Yatsunenko *et al.*, *Nature* **486**, 222 (2012).
14. E. D. Sonnenburg *et al.*, *Nature* **529**, 212 (2016).
15. A. Khoruts *et al.*, *J. Clin. Gastroenterol.* **44**, 354 (2010).

DUAL-USE RESEARCH

Agricultural research, or a new bioweapon system?

Insect-delivered horizontal genetic alteration is concerning

By **R. G. Reeves[1], S. Voeneky[2], D. Caetano-Anollés[1], F. Beck[2], C. Boëte[3]**

Agricultural genetic technologies typically achieve their agronomic aims by introducing laboratory-generated modifications into target species' chromosomes. However, the speed and flexibility of this approach are limited, because modified chromosomes must be vertically inherited from one generation to the next. In an effort to remove this limitation, an ongoing research program funded by the U.S. Defense Advanced Research Projects Agency (DARPA) aims to disperse infectious genetically modified viruses that have been engineered to edit crop chromosomes directly in fields. This is genetic engineering through horizontal transfer, as opposed to vertical inheritance. The regulatory, biological, economic, and societal implications of dispersing such horizontal environmental genetic alteration agents (HEGAAs) into ecosystems are profound. Further, this program stipulates that the means of delivery of these viral HEGAAs into the environment should be insect-based dispersion (1). In the context of the stated aims of the DARPA program, it is our opinion that the knowledge to be gained from this program appears very limited in its capacity to enhance U.S. agriculture or respond to national emergencies (in either the short or long term). Furthermore, there has been an absence of adequate discussion regarding the major practical and regulatory impediments toward realizing the projected agricultural benefits. As a result, the program may be widely perceived as an effort to develop biological agents for hostile purposes and their means of delivery, which—if true—would constitute a breach of the Biological Weapons Convention (BWC).

Announced by DARPA in November 2016 (1), the Insect Allies program is reportedly backed by more than $27 million in awarded research contracts (2–4). In July 2017, the first of three consortia announced that they had been awarded a contract from DARPA to develop systems for insect dispersion of genetically modified viruses (2–4). These are contracts for completion of a 4-year work plan (1) that will culminate in large-scale greenhouse demonstrations of the fully functional insect-dispersed HEGAA approach. Maize and tomato plants are reportedly being used in current experiments (2–4), while dispersal insect species mentioned include leafhoppers, whiteflies, and aphids (3, 5).

Since its public inception, the Insect Al-

> ## "...genetically modified virus to perform gene editing of...crops in...planted fields..."

---

lies program has almost exclusively been presented as a means for farmers to address routine agricultural concerns (e.g., drought, frost, flooding, salinity, herbicides, and disease) (2–4, 6). Despite a 2-year time frame for the generation of a functional proof-of-principle system in greenhouses (1), there has been little public explanation of how developments arising from conducting the work plan of the Insect Allies program might be applied to achieve the agricultural benefits highlighted by the funder or the researchers involved.

HEGAA REGULATORY PATHWAYS?
Although the program work plan does not fully prescribe the nature of the horizontal environmental "genetic alteration agent," it does stipulate the viral expression of ≥3 transgenes to result in a gain-of-function phenotype in crop plants (1). The means by which these gains of function might be achieved remain to be established by the successful applicants, but the simplest route is most likely by editing genes on plant chromosomes (3–6).

Currently, the most obvious candidate for the HEGAAs (1) is a CRISPR system engineered to be part of a virus. Such an approach would target specific plant genes by modifying chromosomes, the alteration of which could increase plant resilience to environmental challenges or herbicides. The end effect of this system would be the use of a genetically modified virus to perform gene editing of susceptible crops in already-planted fields wherever the virus-carrying insects dispersed. It is conceivable that gains of function in the crop could be achieved by the expression of transgenes that were unable to modify the genome of the crop plant. However, numerous statements made by DARPA and involved researchers indicate that a core part of all the consortia's work programs likely includes plant chromosomal editing by means of virally encoded CRISPR proteins (4–6).

Press releases issued by DARPA and the lead institutions of the three consortia (2–4, 6) motivate the research goals exclusively or predominantly by references to routine use in agriculture, e.g., in relation to drought, frost, flooding, herbicides, salinity, or disease. This claim necessarily relies on the assumption that a plausible regulatory pathway exists by which possibly transgenic virus–infected (and potentially genetically altered) food crops could enter national or international markets (4). Even where there was obviously an established regulatory path for such routine use, it would still seem essential to clearly detail it; not to do so would risk undermining the repeatedly stated agricultural aims of the program (2–5). Yet there is no discussion at all in these press releases of regulation—a topic that is central to the question of intent. This omission is all the more striking given that it is likely that all current regulatory systems across the globe would require profound changes to accommodate even an occasional use of HEGAA technology (see the table).

Additionally, it is virtually certain that any concerns expressed by seed producers about how their facilities could be protected would receive considerable attention from regulators, farmers, and governments (see the table). Equally notable is the absence of any public discussion on how measures to ensure the coexistence of different farming practices, often mandated by law, could withstand the use of HEGAA technology.

INSECT DISPERSION AND INTENT
Although the proposition of HEGAA technology is in itself radical, it is DARPA's mandate of using an insect-based means of delivery combined with HEGAAs that sets Insect Allies apart from all other programs. This is because it further increases incompatibilities with current farming practices, as well as the global regulation and trading of bioen-

[1]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, Plön, Germany. [2]Institute of International Law and Ethics of Law, University of Freiburg, Germany. [3]ISEM, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France. Email: reeves@evolbio.mpg.de; svoeneky@jura.uni-freiburg.de

gineered food products, beyond the already substantial challenges stemming from the HEGAA component (see the table).

It would presumably have been possible for the DARPA work program to have proposed the development of HEGAAs to be deployed by using agricultural spraying equipment (7) without the involvement of insects. Crucially, all of the prominently hypothesized benefits to routine peacetime agriculture could probably be realized through the spraying of HEGAAs. It is therefore reasonable to ask, why mandate insect-based dispersion at all? The sole justification that has been put forward in any public document is that overhead spraying of HEGAAs would require infrastructure that is not available to all farmers (1, 6, 8). Although this could conceivably provide a rather weak financial-efficiency justification for farmers in developing nations, it is hardly plausible when applied to the vast majority of crop farmers in the United States.

Although routine agricultural use is prominently presented in most documents as the motivation for the Insect Allies program, a secondary motivation is briefly acknowledged in some—namely its use as a defensive response to unspecified "threats introduced by state or nonstate actors" (6, 8). Only one of the three academic consortia makes clear reference to defensive or emergency applications in its press release (4). However, it is hard to imagine that U.S. farmers would not be ensured access to overhead spraying equipment (most likely via military and civilian aircraft) in the eventuality of such man-made threats. Furthermore, spraying capacity can be much more rapidly scaled-up than the mass production of insects, certainly within the time frame of a crop-growing season. In this respect, it is notable that the vast majority of frontline emergency measures to control insect pests for both agriculture and health continue to rely on spraying, even for pest species where control measures based on the release of live insects have been developed (e.g., sterile males). This is likely a reflection of the difficulty of rapidly scaling-up insect production and distribution systems.

If the DARPA program had proposed to enhance national overhead aerial spraying capacities, rather than stipulate the usage of insects as the means of delivery, it would be hard to argue that this would not result in a much greater enhancement of the United States' capacity to respond to hypothetical emergencies. Consequently, it is our opinion that until DARPA provides suitably robust explanations for the necessity of mandating insect dispersion in routine agricultural or emergency applica-

tions, Insect Allies risks being widely perceived as an attempt to develop a means of delivering HEGAAs for offensive purposes.

BROADER INTERNATIONAL CONTEXT

The importance of perceived intent is of critical consideration in regard to allowing and encouraging scientific research for peaceful purposes while obeying international treaties that prohibit the use and development of certain weapons.

For example, the 1976 UN Convention on Environmental Modification Techniques states that it "shall not hinder the use of environmental modification techniques for peaceful purposes." However, the often-unstated personal or institutional motives of

---

No clear regulatory path

Matters likely to receive scrutiny by regulators and governments, regarding agricultural products from locations to which HEGAAs may have dispersed during crop development or seed production.

1. **HEGAAs with somatic gene editing capability:** Infection-susceptible crop genomes may be modified where the specified chromosomal target is present. Even allowing for future technological developments in gene editing, it is unlikely that all plants in a field will receive the same modification at the intended chromosomal target site. This is quite different to the specific laboratory-generated genetic modifications and associated descriptions of their properties that regulators currently consider.

2. **HEGAAs with germline gene editing capacity:** Introduction of edits into the germplasm of crops could considerably complicate efforts to protect this globally critical resource that, since ancient times, governments have played a fundamental role in securing for the future [see already developed laboratory-confined viral germline editing systems (10,11)].

3. **Insect delivery of HEGAAs:** It will not always be possible to confidently identify which plants or fields have been infected by the genetically modified virus (due to inevitable uncertainty about insect movements and the susceptibility of crops to viral infection). This would be a particularly critical problem in areas where seeds are produced for replanting.

---

research groups and funding bodies cannot be the sole basis upon which the wisdom of research programs is assessed. The 1972 BWC avoids relying solely on intent, insofar as it prohibits the development of biological agents "of types and in quantities that have no justification for prophylactic, protective or other peaceful purposes." Hence, a party engaging in the development of biological agents for which a hostile-use case is plausible (or even obvious, in the case of the Insect Allies program) must present acceptable explanations that its research is only serving peaceful purposes.

In addition, the BWC states that each State Party "undertakes never in any circumstances to develop, produce, stockpile or otherwise acquire or retain...weapons, equipment or means of delivery designed to use such agents or toxins for hostile purposes or in armed conflict." We therefore argue that the clause prohibiting the development of the "means of delivery" is likely to prove of particular relevance to the perceived intent of the Insect Allies program. In 2012, State Parties to the BWC expressly agreed that "certain developments in science and technology have the potential for use contrary to the provisions of the Convention now or in the future. These developments include, inter alia,...to develop novel means of delivering biological agents and toxins." (9)

Whether or not a chromosomal editing system (e.g., CRISPR) is ultimately used to achieve DARPA's stipulated aims, easy simplifications (and not elaborations) of the described work program could be used to generate a new class of biological weapon. For example, it is mandated that released insects must be subject to "conditional lethal safeguard(s)" whereby no released insects survive longer than 2 weeks. This is presumably to limit the dispersal of the HEGAAs (1). However, any program that would simply release insects without taking steps to implement some form of conditional lethality would have the effect of increasing the dispersal of HEGAAs into the environment.

HEGAA weapons could be extremely transmissible to susceptible crop species, particularly where insects were used as the means of delivery. Chromosomal editing would be targetable to particular crop varieties dependent on their genome sequence (presumably those varieties not grown by the deploying parties) (see the figure).

Possibly as a consequence of the program's contract-based funding process, we are unaware of any publicly available assessments of the ethical, trade, biosafety, or international biosecurity implications that would normally accompany such a globally

important program of work (it is likely that congressional committees generated such assessments, but they are not in the public record). As a result, awareness of HEGAAs and insect-based means of their delivery is currently very low among nonspecialist scientists and policy-makers, even though it is anticipated that key development milestones should be achieved within the next year (*1*). Should this be accepted as the global norm for funding projects that enable such potentially hazardous directions of research, the best practices and rules, which have contributed to keeping our world largely free from the use of devastating biological weapons for over 60 years, may be seriously undermined.

acting weapons, along with their means of delivery, capable of threatening virtually any crop species (see the figure).
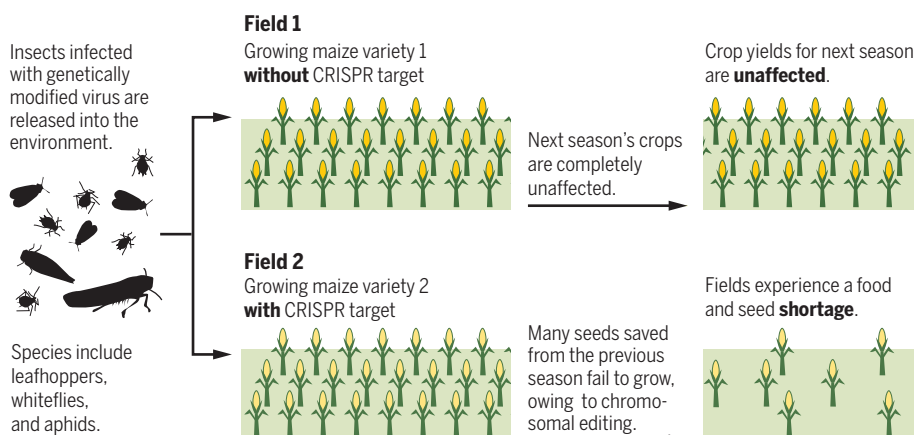
Two of the three consortia funded by DARPA have publicly identified the target species for their experiments as maize— a crop upon which hundreds of millions of people rely for their basic nutritional needs, mainly in Latin America and Africa. Despite the contentious nature of what DARPA was proposing, the agency explicitly discounted funding projects that would target plant species with limited implications for global food security [specifically tobacco or *Arabidopsis* (*1*)]. This reflects the applied nature of DARPA funding, which aims toward making real-world impacts, primarily in the defense

intelligence about another state's objectives with, for example, its recent $100 million programs incorporating Gene Drive (*12*) and Ecological Niche Preference Engineering (*13*). However, the mere announcement of the Insect Allies program, with its presented justifications, may motivate other countries to develop their own capabilities in this arena—indeed, it may have already done so (*14*). This will largely occur independently of whether or not the DARPA program is ultimately scientifically successful, or whether any of the results are made public as planned.

Reversal of funding for this DARPA project by the U.S. Congress [as has happened in the past with DARPA programs that had not received sufficient real-world consideration (*15*)] would not in itself close the particular Pandora's box that HEGAAs or their insect dispersal may represent. Nonetheless, there is a compelling argument that nowhere has bold leadership for the benefit of humankind been more internationally reciprocated than in the control of the use, development, or stockpiling of biological weapons. ∎

A relatively benign hypothetical targeted weaponization scenario involving insect-based delivery of a HEGAA

A viral HEGAA targets chromosomal genes essential for seed fertility in a particular crop variety, the crop species being generally susceptible to genetically modified virus infection. The genetically modified virus is assumed to be capable of infecting either seeds or meristems. For this example, the released virus-infected insects may survive longer than the DARPA stipulated 2 weeks (*1*), or may reacquire virus infection from plants.

Insects infected with genetically modified virus are released into the environment.

Species include leafhoppers, whiteflies, and aphids.

Field 1
Growing maize variety 1 **without** CRISPR target

Next season's crops are completely unaffected.

Crop yields for next season are **unaffected**.

Field 2
Growing maize variety 2 **with** CRISPR target

Many seeds saved from the previous season fail to grow, owing to chromosomal editing.

Fields experience a food and seed **shortage**.

FUTURE INFORMATION HAZARD
There is likely to be widespread agreement with the sentiment that even legitimate biosecurity concerns should not undermine basic research and the development of biotechnologies for the good of humankind (Article X of the BWC). Unfortunately, the extent to which scientific developments will ultimately prove beneficial or harmful is not always straightforward to divine.

To be clear, it is not our contention that the Insect Allies program is ill-conceived simply because it is a military-funded program. Nor would we accept the assertion that the program is less problematic because it has been somewhat transparently initiated with academics. In our view, the program is primarily a bad idea because obvious simplifications of the work plan with already-existing technology (*10*, *11*) can generate predictable and fast-

and security arenas. Although DARPA does "not anticipate applying publication restrictions" to results generated by the three research consortia [section 2.2 of (*1*)], one could argue that this may need to be reconsidered to avoid the proliferation of what may be seen as preliminary instruction manuals on how to develop offensive HEGAA programs, directed in the first instance against maize.

BLURRED LINES
It is worth restating that weapons programs are often driven by perceptions of competitors' activities. Thus, it can be argued that participants in programs that operate close to the blurred lines dividing peacetime and wartime applications are obligated to project robust and plausible motivations for their work. It could, of course, be the case that DARPA is responding to specific

REFERENCES AND NOTES
1. Broad Agency Announcement Insect Allies, Biological Technologies Office, HR001117S0002, 1 November 2016; www.fbo.gov/utils/view?id=40638c9e7d45ed8310f9d4f4671b4a7b.
2. BTI receives DARPA "Insect Allies" Award. *EurekAlert!* (2017); www.eurekalert.org/pub_releases/2017-07/bti-brd072717.php.
3. Penn State team receives $7M award to enlist insects as allies for food security (2017); http://news.psu.edu/story/495037/2017/11/20/research/penn-state-team-receives-7m-award-enlist-insects-allies-food.
4. Ohio State scientists make plant virus system "turn on its head" with insect research (2017); www.thelantern.com/2017/12/ohio-state-scientists-to-make-plant-virus-system-turn-on-its-head-with-insect-research/.
5. Insect Allies: How the Enemies of Corn May Someday Save It (2017); https://cals.ncsu.edu/news/insect-allies-how-the-enemies-of-corn-may-someday-save-it/.
6. DARPA Enlists Insects to Protect Agricultural Food Supply & Commodity Crops. *R&D Mag.* (2016); www.rdmag.com/news/2016/10/darpa-enlists-insects-protect-agricultural-food-supply-commodity-crops.
7. J. S. Cory *et al.*, *Nature* **370**, 138 (1994).
8. Dr. Blake Bextine, Insect Allies; www.darpa.mil/program/insect-allies.
9. Report of the Meeting of the States Parties to the Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on Their Destruction, Geneva (2012); http://undocs.org/en/BWC/MSP/2012/5.
10. Z. Ali *et al.*, *Mol. Plant.* **8**, 1288 (2015).
11. K. Musiychuk *et al.*, *Influenza Other Respir. Viruses.* **1**, 19 (2007).
12. DARPA: Safe Genes; www.darpa.mil/program/safe-genes.
13. DARPA: Young Faculty Awards 2017; www.darpa.mil/work-with-us/for-universities/young-faculty-award.
14. United States of America, Strengthening of the BWC, Meeting of the States Parties to the Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on Their Destruction, Geneva (2018); http://undocs.org/en/BWC/MSP/2018/MX.5/WP.3.
15. J. W. Schoen, Pentagon kills 'terror futures market.' *msnbc.com* (2003); www.nbcnews.com/id/3072985/t/pentagon-kills-terror-futures-market/.

BOOKS *et al.*

FOOD SCIENCE

Junk food, junk science?

A nutrition expert aims a critical eye at the research and marketing practices of food companies

By **Cyan James**

Nutrition expert Marion Nestle wasn't expecting her name to come up in the Democratic Party emails leaked by anonymous hackers during the 2016 presidential campaign. Yet that is exactly what happened. A Clinton campaign adviser, it turns out, was also consulting for Coca-Cola, and the leaked emails included a message from an Australian public relations agency advising the company to keep a watchful eye on Nestle and any concerning statements she might issue.

It turns out that Nestle is perfectly capable of returning the favor. In her latest book, *Unsavory Truth*, Nestle levels a withering fusillade of criticism against food and beverage companies that use questionable science and marketing to push their own agendas about what should end up on our dinner tables.

In her comprehensive review of companies' nutrition research practices, Nestle

Unsavory Truth
Marion Nestle
Basic Books, 2018.
320 pp.

(whose name is pronounced "Nes-sul" and who is not affiliated with the Swiss food company) reveals a passion for all things food, combined with pronounced disdain for the systematic way food companies influence the scientists and research behind the contents of our pantries. Her book is an autopsy of modern food science and advertising, pulling the sheet back on scores of suspect practices, as well as a chronicle of Nestle's own brushes with food company courtship.

Is it shocking that many food companies do whatever they can in the name of fatter profits? Maybe, but it's old hat for Nestle, who has spent five decades honing her expertise and is a leading scholar in the field of nutrition science. In this book, she details nearly every questionable food company tactic in the playbook, from companies that fund their own food science research centers and funnel media attention to nondietary explanations for obesity, to those that cherry-pick data or fund professional conferences as a plea for tacit approval.

Even companies that hawk "benign" foods such as blueberries, pomegranate juice, and nuts come under the author's strict scrutiny

Even the health benefits of seemingly benign foods such as fruits, nuts, and seeds are frequently exaggerated.

because, as she reminds readers, "Foods are not drugs. To ask whether one single food has special health benefits defies common sense."

Instead, Nestle urges eaters to look behind the claims to discover who funds food-science studies, influences governmental regulations, advises policy-makers, and potentially compromises researchers. Food fads, we learn, can spring from a few findings lifted out of context or interpreted with willful optimism. Companies, after all, can hold different standards for conducting and interpreting research than those of independent academic institutions or scientific organizations and can be driven by much different motives.

Nestle wields a particularly sharp pen against junk-food purveyors, soda companies, and others who work hard to portray excess sugar as innocuous or who downplay the adverse effects their products can have. These companies adopt tactics such as founding research centers favorable to their bottom line, diverting consumer attention away from diet to focus instead on the role of exercise in health, or trying to win over individual researchers to favorably represent their products. Nestle calls foul on these attempts to portray junk foods as relatively harmless, even knocking the "health halo" shine from dark chocolates and other candies masquerading as responsible health foods.

Not content to point to the many thorny problems lurking behind food company labels and glitzy sponsored meetings, however, Nestle offers a constructive set of suggestions for how nutrition scientists can navigate potential conflicts of interest.

Consumers, too, bear a responsibility for promoting eating habits that steer clear of dubious advertising. Nestle advocates adhering to a few simple guidelines: "[E]at your veggies, choose relatively unprocessed foods, keep junk foods to a minimum, and watch out for excessive calories." We have the power of our votes and our forks, she reminds us, and can use both to insist that food companies help us eat well. Nestle's determination to go to bat for public health shines through, illuminating even her occasional sections of workaday prose or dense details.

Nestle marshals a convincing number of observations on modern food research practices while energetically delineating how food companies' clout can threaten the integrity of the research performed on their products. There is indeed something rotten in the state of dietary science, but books like this show us that we consumers also hold a great deal of power. ∎

10.1126/science.aau6602

The reviewer is at Peraton, Herndon, VA 20170, USA.
Email: cyanj@uw.edu

COMPUTER SCIENCE

Mythical androids and ancient automatons

Technology tales from classical literature reveal the storied history of artificial intelligence

By **Sarah Olson**

Long before the advent of modern robots and artificial intelligence (AI), automated technology existed in the storytelling and imaginations of ancient societies. Hephaestus, blacksmith of the mythical Greek gods, fashioned mechanical serving maids from gold and endowed them with learning, reason, and skill in Homer's *Iliad*. Designed to anticipate their master's requests and act on them without instruction, the Golden Maidens share similarities with modern machine learning, which allows AI to learn from experience without being explicitly programmed.

The possibility of AI servants continues to tantalize the modern imagination, resulting in household automatons such as the Roomba robot vacuum and Amazon's virtual assistant, Alexa. One could imagine a future iteration of Alexa as a fully automatic robot, perhaps washing our dishes or reminding us to pick up the kids after school. In her new book, *Gods and Robots*, Adrienne Mayor draws comparisons between mythical androids and ancient robots and the AI of today.

Mayor is not the first expert to show how humanity's ancient fascination with automatons can be seen throughout the stories and myths of classical literature. Historians of robotics point to unprecedented techniques—both imagined and innovated—that existed throughout antiquity. Most fascinating, Mayor notes, is the predictive nature of these devices. Odysseus's pilotless Phaeacian ships, for example, which guided him back home to Ithaca, are reminiscent of modern GPS technology.

Through detailed storytelling and careful analysis of popular myths, Mayor urges readers to consider lessons learned from these stories as we set about creating a new world with AI. Like Pandora, an automaton created by Hephaestus who opened a box and released evils into the world, AI could bring about

The reviewer is at MiraCosta College, Oceanside, CA 92056, USA. She reviews books on her website, readmorescience.com. Email: sarahelisabetholson@gmail.com

unforeseen problems. Here, Mayor cites Microsoft's 2016 experiment with the Twitter chatbot "Tay," a neural network programmed to converse with Twitter users without requiring supervision. Only hours after going live, Tay succumbed to a group of followers who conspired to turn the chatbot into an internet troll. Microsoft's iteration the following year suffered a similar fate.

Despite her extensive knowledge of ancient mythology, Mayor does little to demonstrate an understanding of modern AI,


In the absence of proper foresight, artificial intelligence, like Pandora, could unleash unforeseen evils into the world.

neural networks, and machine learning; the chatbots are among only a handful of examples of modern technology she explores.

Instead, Mayor focuses on her own area of expertise: analyzing classical mythology. She recounts, for example, the story of the bronze robot Talos, an animated statue tasked with defending the Greek island of Crete from pirates. Here, she mentions a video game called "The Talos Principle," in

Gods and Robots
Myths, Machines, and Ancient Dreams of Technology
Adrienne Mayor
Princeton University Press, 2018. 291 pp.
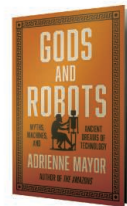
which players assume the role of a robot and solve philosophical puzzles, but she only briefly summarizes it. Because Mayor's analyses are so superb, further exploration into the influence of such stories on makers of modern video games would likely have fascinated readers.

A few sentences to explain the difference between, say, machine learning and AI would have likewise been worthwhile. Such omissions make it difficult to identify the book's intended audience. Although Mayor's comparisons between modern technology and ancient myth might be of some interest to experts on AI, reading will require a strong curiosity for, if not previous knowledge of, classical literature and mythology.

Mayor often repeats herself in her writing—for example, offering multiple defenses for why an ancient creation may be considered an automaton. This repetition can feel as though she doesn't trust readers to follow along with the connections she makes between technology and ancient myth. Yet instead of expanding on several of her more interesting points—the notion that ancient people believed that there were ways to create artificial life without the use of magic, for example—she tends to reference citations in the appendix. Such a dry approach falls short alongside her otherwise wonderful storytelling, thorough research, and impressive expertise.

When Pandora opens her box of evils, she releases hope by accident—inadvertently helping humanity learn to live in a world now corrupted by evil. Like this story, *Gods and Robots* is cautionary but optimistic. While warning of the risks that accompany the irresponsible deployment of technology, Mayor reassures readers that AI could indeed bring about many of the wonderful things that our ancestors imagined. ∎
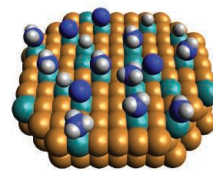
10.1126/science.aav0325

Light-dependent activation in photocatalysis

Zhou et al., p. 69

Edited by **Stella Hurtley**

Boulder-filled channel near the eastern margin of the Tibetan Plateau, a setting similar to the Bhotekoshi River

GEOMORPHOLOGY
A sudden outburst of erosion

Glacial lake outburst floods (GLOFs) are exactly what they sound like. The sudden emptying of a glacial lake in high-topography regions like the Himalaya can quickly destroy everything in its path. Cook *et al*. intercepted a GLOF in the Bhotekoshi and Sunkoshi river valleys in central Nepal as they were monitoring the region in the aftermath of the 2015 Gorkha earthquake. They found that a massive amount of erosion occurred during the outburst flood, which suggests that GLOFs may be the primary factor in landscape evolution for these regions. —BG

Science, this issue p. 53

SUPERCONDUCTIVITY
Conspiring interactions in a cuprate

More than 30 years after the discovery of high-temperature superconductivity in copper oxides, its mechanism remains a mystery. Electron pairing mediated solely by lattice vibrations—phonons—is thought to be insufficient to account for the high transition temperatures. He *et al*. found a rapid and correlated increase of the superconducting gap and electron-phonon interactions as the chemical composition of their bismuth-based cuprate samples was varied across a critical doping concentration. The interplay of electron-phonon with electron-electron interactions may lead to enhanced transition temperatures. —JS

Science, this issue p. 62

ADAPTIVE RADIATIONS
Secrets revealed by kangaroo teeth

The teeth of mammals display complex adaptations to diet and can thus provide a window into the environments of extinct species. Couzens and Prideaux used such a window to examine the expansion and diversification of kangaroos, Australia's largest herbivores (see the Perspective by Polly). True kangaroos diversified not in response to drying in the Miocene, as suggested by molecular results, but rather as grasslands expanded during the Pliocene. Furthermore, the now-extinct short-faced kangaroos were not declining because of increases in aridity at the end of the Pleistocene but instead were experiencing an increase in dietary divergence. —SNV

Science, this issue p. 72; see also p. 25

INFLUENZA
Seasonal flu by ZIP code

Influenza virus strikes communities in northern latitudes during winter, straining health care provision almost to the breaking point. Change in environmental humidity is a key driver, but many other seasonal and social factors contribute. Dalziel *et al*. obtained a geographical distribution of doctor visits for influenza-like illness for more than 600 U.S. cities (see the Perspective by Wallinga). Some ZIP codes regularly experienced sharply defined peaks of cases, or intense epidemics, and others showed a longer, more diffuse influenza season. The surges tended to occur in smaller cities with less residential density and lower household incomes. Larger, more densely populated cities had more-diffuse epidemics, presumably because of higher rates of personal

contact, which makes influenza transmission less subject to climate variation. —CA

Science, this issue p. 75;
see also p. 29

ECONOMICS
Educating for economic rationality

The hypothesis that education enhances economic decision-making has been surprisingly underexplored. Kim *et al*. studied this question using a randomized control trial in a sample of 2812 girls in secondary schools in Malawi. Four years after providing financial support for a year's schooling, they presented the subjects with a set of decision problems (for example, allocating funds to immediate versus future expenses) that test economic rationality. The education intervention enhanced both educational outcomes and economic rationality as measured by consistency with utility maximization in the long run. —AMS

Science, this issue p. 83

MUSCLE DISEASE
Gene editing and muscular dystrophy

Duchenne muscular dystrophy (DMD) is characterized by progressive muscle weakness and a shortened life span. The disease is caused by mutations that reduce or prevent expression of dystrophin, an essential structural protein in skeletal and heart muscle. The gene editing technology CRISPR-Cas9 can correct disease-causing mutations and has yielded promising results in mouse models of DMD. In a small, short-term study, Amoasii *et al*. tested this strategy in a dog model of DMD that exhibits many features of the human disease. Intramuscular or systemic delivery of the gene editing components resulted in a substantial increase in dystrophin protein levels in skeletal and heart muscle. Restoration

of dystrophin expression was accompanied by improved muscle histology. —PAK

Science, this issue p. 86

ASTRONOMY
New moon rising

Although the existence of exomoons—moons orbiting extrasolar planets—is probable, direct observational evidence has been elusive. Previous observations made using NASA's Kepler space telescope suggested that Kepler-1625b, a Jupiter-sized planet orbiting the solar-mass yellow star Kepler-1625, may be orbited by an exomoon. Now, Teachey and Kipping report additional observations made using the Hubble Space Telescope, as well as a refined analysis of Kepler photometry data, that strongly support the exomoon hypothesis. This moon, if it exists, would be similar in size to Neptune or Uranus in our own Solar System. —KVH

Sci. Adv. 10.1126/sciadv.aav1784 (2018).

HIV
Gut check for a promising HIV treatment

Eradicating HIV in infected patients likely requires disrupting the reservoir of infected T cells in the gastrointestinal tract. One approach may be targeting cells expressing the integrin α4β7, which has been tested in simian immunodeficiency virus models and is an approved therapy for inflammatory bowel disease. Uzzan *et al*. studied a small cohort of HIV-infected individuals on antiretroviral therapy who began receiving an antibody against α4β7 as a treatment for their mild inflammatory bowel disease. Longitudinal colonoscopies revealed that the anti-α4β7 therapy disrupted local lymphoid aggregates. The treatment was well tolerated, but long-term effects on the HIV reservoir remain to be determined. —LP

Sci. Transl. Med. **10**, eaau4711 (2018).

Osteoporosis reduces bone density because of imbalances in signaling feedbacks.

GENOMICS
Evolving risks to personality

The origin of the uniquely large and complex human brain has been shrouded in mystery. To identify genetic changes that may underlie human evolution, Sato and Kawata examined human-specific genetic changes associated with neuropsychiatric disease and identified selection for variants in three genes. Interestingly, one variant of the *SLC18A1* gene, which encodes a neurotransmitter, arose at the time humans migrated out of Africa. It appears to be under balancing selection in some populations, which has resulted in two variants in most non-African populations. This variant is linked with anxiety, but it is not yet understood what conditions favor its selection. —LMZ

Evol. Lett. **2**, 499 (2018).

SOCIOLOGY
Joke theft and comedians' authenticity

Joke theft among comedians is difficult to discern because jokes are rarely identical and may be discovered independently of each other. To determine how people adjudicate instances of joke theft, Reilly conducted 4 years of participant observation and interviews among stand-up comics in Los Angeles to examine norms surrounding joke theft. He found that accusations of joke theft are driven less by joke similarity and more by a comedian's reputation, authenticity, and past success. Comedians perceived as being more successful than they deserve to be are more vulnerable to accusations of joke theft. These data have implications for understanding how intellectual property disputes are handled in other creative and scientific domains, as well as how character and action information are

PHOTO: ALAN BOYDE/VISUALS UNLIMITED, INC.
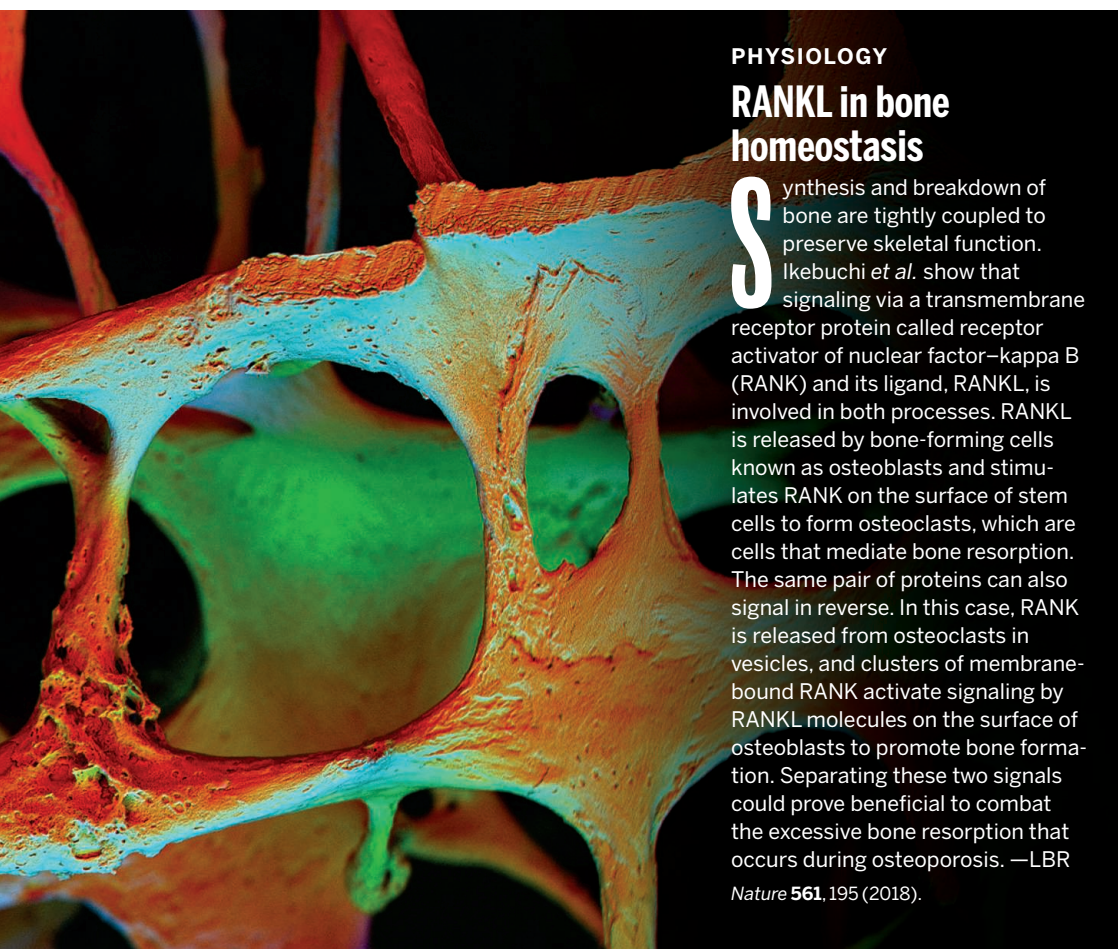
PHYSIOLOGY
RANKL in bone homeostasis

Synthesis and breakdown of bone are tightly coupled to preserve skeletal function. Ikebuchi *et al.* show that signaling via a transmembrane receptor protein called receptor activator of nuclear factor–kappa B (RANK) and its ligand, RANKL, is involved in both processes. RANKL is released by bone-forming cells known as osteoblasts and stimulates RANK on the surface of stem cells to form osteoclasts, which are cells that mediate bone resorption. The same pair of proteins can also signal in reverse. In this case, RANK is released from osteoclasts in vesicles, and clusters of membrane-bound RANK activate signaling by RANKL molecules on the surface of osteoblasts to promote bone formation. Separating these two signals could prove beneficial to combat the excessive bone resorption that occurs during osteoporosis. —LBR

Nature **561**, 195 (2018).

dendrite order might facilitate nonsynaptic signaling across dendrites. —PJH

eLife **7**, e35825 (2018).

THERMAL TRANSPORT
Blowing by the blackbody limit

The blackbody limit tells us the maximum amount of heat that can be emitted from an object. However, this limit is known to be no limit at all when the spacing between objects is small enough. Thompson *et al.* show that the blackbody limit can be circumvented when the emitting objects themselves are small. More importantly, they demonstrate a massive, hundredfold enhancement relative to the classical blackbody limit using nanoscale membranes. This opens the door to tailoring nanoscale materials for enhanced thermal absorption and emission applications. —BG

Nature **561**, 216 (2018).

ELECTRON IMAGING
Imaging with ghostly electrons

Ghost imaging is a computational imaging approach that uses correlations between a signal beam that interacts with a sample and a reference beam that doesn't. The correlations are then used to reconstruct an image of the sample and can, in principle, be used to image samples with very weak signal beams that might otherwise be damaged with a typical probe beam. Already demonstrated for optics and atoms, Li *et al.* now extend the method of ghost imaging to electrons. To circumvent the difficulty of splitting an electron beam, they instead used a structured electron beam to probe the sample. They show that a reconstructed image of a sample with a lower electron dose than that used in direct imaging methods can be faithfully produced. This technique could reduce acquisition time and avoid damaging sensitive samples in electron imaging applications. —ISO

Phys. Rev. Lett. **121**, 114801 (2018).

integrated into moral judgment more broadly. —TSR

Am. Sociol. Rev. **83**, 933 (2018).

HEALTH
Gene for preeclampsia

Preeclampsia is a serious complication of pregnancy that often presents as an increase in maternal blood pressure. The disorder is more prevalent and severe in women with African ancestry, most likely because of genetic factors. Reidy *et al.* have been investigating a variant of the gene encoding apolipoprotein L1 (*APOL1*), which was previously shown to confer a high risk of kidney disease in black Americans. Studying two independent populations of pregnant black women, they found that preeclampsia was associated with the *APOL1* high-risk genotype. Interestingly, it was the genotype of the fetus, not the mother, that mattered. The fetal *APOL1* high-risk genotype may be linked to one in eight cases of preeclampsia in black women. —PAK

Am. J. Hum. Genet. **103**, 367 (2018).

NEURODEVELOPMENT
Sticky order for neurites

How do biological systems assemble in an ordered way, particularly in complex organs like the brain? One idea is that order emerges from multiple simple, local interactions. To explore this question, Yip and Heiman examined the organization of the sensory neurons of the nematode *Caenorhabditis elegans*. The amphid sense organ sends processes to the anterior of the worm to detect environmental conditions. The dendrites that make up these nerves are bundled in an ordered manner. Close to the nose, one dendrite occupies the middle of the bundle. A little further away from the nose, this dendrite cedes the middle track to another dendrite. The ordering is determined by cell-adhesion molecules expressed by the neurons themselves. The stereotypical organization of
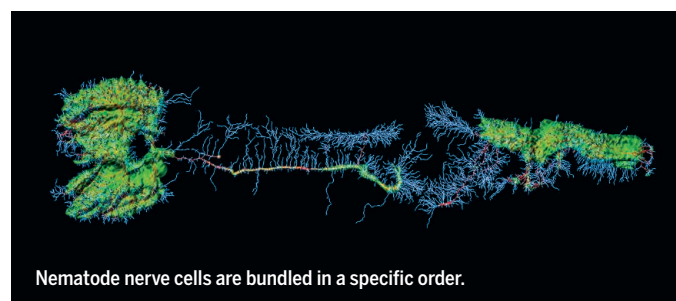
Nematode nerve cells are bundled in a specific order.

ALSO IN *SCIENCE* JOURNALS

Edited by **Stella Hurtley**

GAS GIANT PLANETS
Cassini's final phase of exploration

The Cassini spacecraft spent 13 years orbiting Saturn; as it ran low on fuel, the trajectory was changed to sample regions it had not yet visited. A series of orbits close to the rings was followed by a Grand Finale orbit, which took the spacecraft through the gap between Saturn and its rings before the spacecraft was destroyed when it entered the planet's upper atmosphere. Six papers in this issue report results from these final phases of the Cassini mission. Dougherty *et al.* measured the magnetic field close to Saturn, which implies a complex multilayer dynamo process inside the planet. Roussos *et al.* detected an additional radiation belt trapped within the rings, sustained by the radioactive decay of free neutrons. Lamy *et al.* present plasma measurements taken as Cassini flew through regions emitting kilometric radiation, connected to the planet's aurorae. Hsu *et al.* determined the composition of large, solid dust particles falling from the rings into the planet, whereas Mitchell *et al.* investigated the smaller dust nanograins and show how they interact with the planet's upper atmosphere. Finally, Waite *et al.* identified molecules in the infalling material and directly measured the composition of Saturn's atmosphere. —KTS

Science, this issue p. 46, p. 47, p. 48, p. 49, p. 50, p. 51

STRUCTURAL BIOLOGY
It takes two to signal

The Hedgehog (HH) signaling pathway is important in development, and excessive HH signaling is associated with cancer. Signaling occurs through the G protein–coupled receptor Smoothened. The pathway is repressed by the membrane receptor Patched-1 (PTCH1),

and this inhibition is relieved when PTCH1 binds the secreted protein HH. Two recent papers have described structures of HH bound to PTCH1, but surprisingly, each described a different binding epitope on HH. Qi *et al.* present a cryo–electron microscopy structure that explains this apparent contradiction by showing that a single HH protein uses both of these interfaces to engage two PTCH1 receptors (see the Perspective by Sommer and Lemmon). Functional assays suggest that both interfaces must be bound for efficient signaling. —VV

Science, this issue p. 52; see also p. 26

GEOPHYSICS
Detailing subduction zones

Subduction zones are responsible for the most-damaging and tsunami-generating great earthquakes. Hayes *et al.* updated their Slab1.0 model to include all seismically active subduction zones, including geometrically complex regions like the Philippines. The new model, Slab2, details the geometry of 24 million square kilometers of subducted slabs, from ocean trench to upper mantle. The model will be vital for fully understanding seismic hazard in some of the most populated regions in the world. —BG

Science, this issue p. 58

SOLID-STATE PHYSICS
Insulator or a metal?

When a metal is cooled to low temperatures and placed in an external magnetic field, its resistivity may oscillate as the magnitude of the field is varied. Seeing these so-called quantum oscillations in an insulating material would be very unusual. Xiang *et al.* report such findings in the insulator ytterbium dodecaboride (YbB_{12}) (see the Perspective by Ong). In addition

to oscillations in resistivity, the authors observed oscillations in the magnetic torque. The results present a challenge to theories that aim to explain the insulating state of YbB_{12}. —JS

Science, this issue p. 65; see also p. 32

PLASMONIC CHEMISTRY
Hot carriers reducing thermal barriers

Plasmonic catalysts can generate hot charge carriers that can activate reactants and, in turn, reduce the overall barrier to a reaction. Zhou *et al.* studied the decomposition of ammonia to hydrogen on a copper alloy nanostructure that absorbed light and generated electrons that activated nitrogen atoms on ruthenium surface atoms (see the Perspective by Cortés). By measuring reaction rates at different wavelengths, light intensities, and catalyst surface temperatures, the light-induced reduction of the apparent activation barrier was quantified. —PDS

Science, this issue p. 69; see also p. 28

FOREST ECOLOGY
Tree diversity improves forest productivity

Experimental studies in grasslands have shown that the loss of species has negative consequences for ecosystem functioning. Is the same true for forests? Huang *et al.* report the first results from a large biodiversity experiment in a subtropical forest in China. The study combines many replicates, realistic tree densities, and large plot sizes with a wide range of species richness levels. After 8 years of the experiment, the findings suggest strong positive effects of tree diversity on forest productivity and carbon accumulation. Thus, changing from monocultures to more mixed forests could benefit both restoration

of biodiversity and mitigation of climate change. —AMS

Science, this issue p. 80

CANCER
Some (re)programming notes on cancer

Epithelial cancers develop resistance to targeted therapies in a number of different ways. Several cancer types do so by undergoing phenotypic conversion to a highly aggressive cancer called small cell neuroendocrine carcinoma (SCNC). Whether distinct cancer types accomplish this "reprogramming" through the same mechanism has been unclear. Park *et al.* show that the same set of oncogenic factors transforms both normal lung and normal prostate epithelial cells into SCNCs that resemble clinical samples (see the Perspective by Kareta and Sage). This convergence of molecular pathways could potentially simplify the development of new therapies for SCNC, which is currently untreatable. —PAK

Science, this issue p. 91; see also p. 30

MICROBIOLOGY
Industrialization reduces microbiota diversity

Disturbances in the microbiota that live on or in the human body are associated with various diseases. In a Perspective, Dominguez Bello *et al.* propose that a reduction in microbiota diversity associated with modern, industrialized living is linked to the increasing incidence of metabolic, immune, and cognitive diseases. They propose that microbiota samples should be collected from individuals in traditional societies. These could potentially be used to reinstate microbiota ecology, which might help to prevent and/or treat modern diseases. —GKA

Science, this issue p. 33

Turning the tables on interferon

An early step in the host response to viral infection involves a burst of synthesis of type I interferons that allow cells to quickly fight back against the offending viruses. Shaabani *et al*. investigated how the same interferon-stimulated genes (ISGs) that usually help against viruses surprisingly dampen the host's ability to resist many bacterial infections. Deletion of a single ISG called *Usp18* in mouse dendritic cells was sufficient to enhance host control of infections with two strains of Gram-positive bacteria. Normal induction of USP18 after infection impaired antibacterial responses mediated by tumor necrosis factor. USP18 thus represents a potential therapeutic target for control of serious bacterial infections. —IW

Sci. Immunol. **3**, eaau2125 (2018).

How *Vibrio* disrupts Ras signaling

Many pathogenic bacteria, including *Vibrio vulnificus*, target the small guanosine triphosphatase Ras because it is critical for signaling pathways that control host cell biology and innate defenses. Biancucci *et al*. solved the crystal structure of RRSP, a toxin effector domain from *V. vulnificus* that cleaves KRAS, a member of the Ras family. Cleavage by RRSP did not release any fragments but structurally altered KRAS so that it could not bind to its downstream effector Raf. These findings may prove useful for developing strategies to inhibit Ras proteins that are aberrantly activated in tumors. —AV

Sci. Signal. **11**, eaat8335 (2018).

DIVING WITHIN
SATURN'S RINGS

By **Keith T. Smith**

L aunched in 1997, the Cassini spacecraft arrived at Saturn in 2004 and spent the following 12 years orbiting far outside the planet's rings. Among the mission's many discoveries were potentially habitable environments on Saturn's moons, including a subsurface ocean of water on Enceladus.
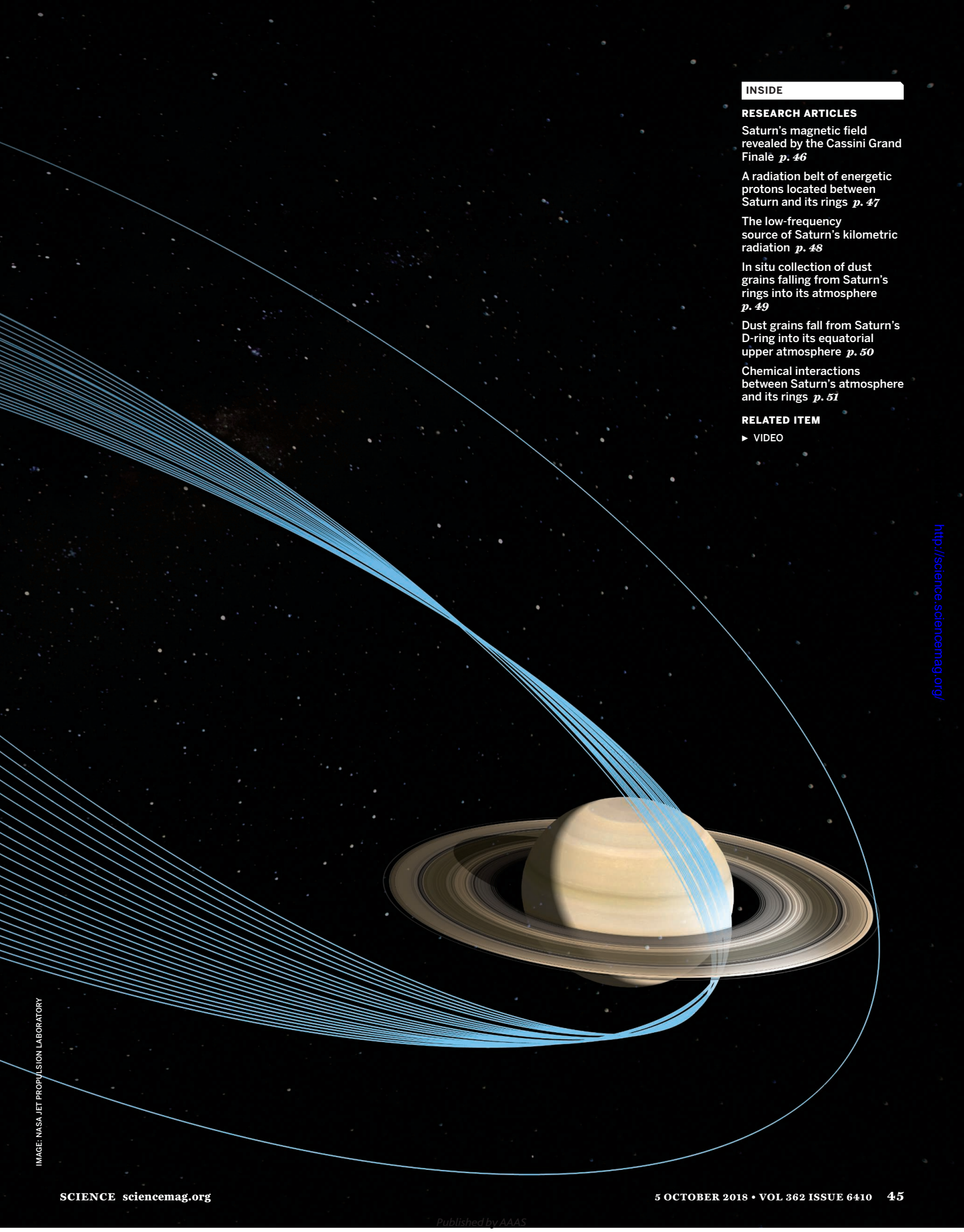
To avoid accidental contamination of the moons with any Earth bacteria that might have hitched a ride on the spacecraft, mission planners decided that when Cassini ran out of fuel it would be deliberately crashed into Saturn. As fuel reserves ran low, the spacecraft embarked on a series of maneuvers that took it through previously unexplored regions and ended with its destruction.

The first phase was a series of 20 "ring-grazing" orbits in which Cassini passed near the outer edge of the rings. Then the trajectory was shifted to fly through the gap between the planet and the inner edge of the rings, a phase dubbed the "Grand Finale." Cassini flew through this region 22 times, in various orientations designed to optimize the scientific measurements made by its numerous instruments. Finally, in 2017 Cassini dove into the planet itself at 35 km s^{-1}, leading to the spacecraft's fiery disintegration in the upper atmosphere.

The papers in this section present results from those final phases of the mission. They include studies of Saturn's puzzling magnetic field, an isolated inner radiation belt, auroral radio emission, the mass and composition of solid grains falling from the rings toward the planet, and the composition of Saturn's atmosphere. Many more results will follow from Cassini's last period of exploration in the Saturn system.

A diagram of Cassini's final orbits (blue curves), showing the last ring-grazing orbit and the 22 Grand Finale orbits.

http://science.sciencemag.org/

RESEARCH ARTICLE SUMMARY

GAS GIANT PLANETS

Saturn's magnetic field revealed by the Cassini Grand Finale

Michele K. Dougherty*, Hao Cao, Krishan K. Khurana, Gregory J. Hunt, Gabrielle Provan, Stephen Kellock, Marcia E. Burton, Thomas A. Burk, Emma J. Bunce, Stanley W. H. Cowley, Margaret G. Kivelson, Christopher T. Russell, David J. Southwood

INTRODUCTION: Starting on 26 April 2017, the Grand Finale phase of the Cassini mission took the spacecraft through the gap between Saturn's atmosphere and the inner edge of its innermost ring (the D-ring) 22 times, ending with a final plunge into the atmosphere on 15 September 2017. This phase offered an opportunity to investigate Saturn's internal magnetic field and the electromagnetic environment between the planet and its rings. The internal magnetic field is a diagnostic of interior structure, dynamics, and evolution of the host planet. Rotating convective motion in the highly electrically conducting layer of the planet is thought to maintain the magnetic field through the magnetohydrodynamic (MHD) dynamo process. Saturn's internal magnetic field is puzzling because of its high symmetry relative to the spin axis, known since the Pioneer 11 flyby. This symmetry prevents an accurate determination of the rotation rate of Saturn's deep interior and challenges our understanding of the MHD dynamo process because Cowling's theorem precludes a perfectly axisymmetric magnetic field being maintained through an active dynamo.

RATIONALE: The Cassini fluxgate magnetometer was capable of measuring the magnetic field with a time resolution of 32 vectors per s and up to 44,000 nT, which is about twice the peak field strength encountered during the Grand Finale orbits. The combination of star cameras and gyroscopes onboard Cassini provided the attitude determination required to infer the vector components of the magnetic field. External fields from currents in the magnetosphere were modeled explicitly, orbit by orbit.

RESULTS: Saturn's magnetic equator, where the magnetic field becomes parallel to the spin axis, is shifted northward from the planetary equator by 2808.5 ± 12 km, confirming the north-south asymmetric nature of Saturn's magnetic field. After removing the systematic variation with distance from the spin axis, the peak-to-peak "longitudinal" variation in Saturn's magnetic equator position is <18 km, indicating that the magnetic axis is aligned with the spin axis to within 0.01°. Although structureless in the longitudinal direction, Saturn's internal magnetic field features variations in the latitudinal direction across many different characteristic length-scales. When expressed in spherical harmonic space, internal axisymmetric magnetic moments of at least degree 9 are needed to describe the latitudinal structures. Because there was incomplete latitudinal coverage during the Grand Finale orbits, which can lead to nonuniqueness in the solution, regularized inversion techniques were used to construct an internal Saturn magnetic field model up to spherical harmonic degree 11. This model matches Cassini measurements and retains minimal internal magnetic energy. An azimuthal field component two orders of magnitude smaller than the radial and meridional components is measured on all periapses (closest approaches to Saturn). The steep slope in this component and magnetic mapping to the inner edge of the D-ring suggests an external origin of this component.

CONCLUSION: Cassini Grand Finale observations confirm an extreme level of axisymmetry of Saturn's internal magnetic field. This implies the presence of strong zonal flows (differential rotation) and stable stratification surrounding Saturn's deep dynamo. The rapid latitudinal variations in the field suggest a second shallow dynamo maintained by the background field from the deep dynamo, small-scale helical motion, and deep zonal flows in the semiconducting region closer to the surface. Some of the high-degree magnetic moments could result from strong high-latitude concentrations of magnetic flux within the planet's deep dynamo. The periapse azimuthal field originates from a strong interhemispherical electric current system flowing along magnetic field lines between Saturn and the inner edge of the D-ring, with strength comparable to that of the high-latitude field-aligned currents (FACs) associated with Saturn's aurorae. ■
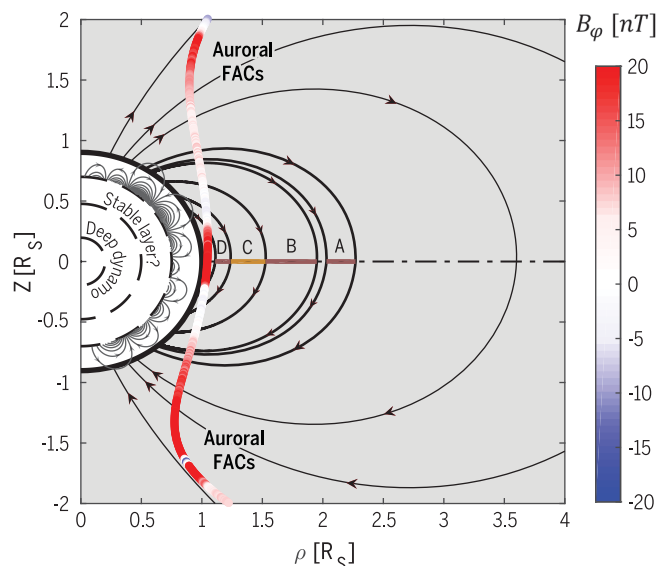
A meridional view of the results of the Cassini magnetometer observations during the Grand Finale orbits. Overlain on the spacecraft trajectory is the measured azimuthal field from the first Grand Finale orbit, revealing high-latitude auroral FACs and a low-latitude interhemispherical FAC system. Consistent small-scale axisymmetric internal magnetic field structures originating in the shallow interior are shown as field lines within the planet. A tentative deep stable layer and a deeper dynamo layer, overlying a central core, are shown as dashed semicircles. The A-, B-, C-, and D-rings are labeled, and the magnetic field lines are shown as solid lines. R_S is Saturn's radius, Z is the distance from the planetary equator, ρ is the perpendicular distance from the spin axis, and B_ϕ is the azimuthal component of the magnetic field.

RESEARCH ARTICLE

GAS GIANT PLANETS

Saturn's magnetic field revealed by the Cassini Grand Finale

Michele K. Dougherty[1]*, Hao Cao[2,3], Krishan K. Khurana[4], Gregory J. Hunt[1],
Gabrielle Provan[5], Stephen Kellock[1], Marcia E. Burton[6], Thomas A. Burk[6],
Emma J. Bunce[5], Stanley W. H. Cowley[5], Margaret G. Kivelson[4,7],
Christopher T. Russell[4], David J. Southwood[1]

During 2017, the Cassini fluxgate magnetometer made in situ measurements of Saturn's magnetic field at distances ~2550 ± 1290 kilometers above the 1-bar surface during 22 highly inclined Grand Finale orbits. These observations refine the extreme axisymmetry of Saturn's internal magnetic field and show displacement of the magnetic equator northward from the planet's physical equator. Persistent small-scale magnetic structures, corresponding to high-degree (>3) axisymmetric magnetic moments, were observed. This suggests secondary shallow dynamo action in the semiconducting region of Saturn's interior. Some high-degree magnetic moments could arise from strong high-latitude concentrations of magnetic flux within the planet's deep dynamo. A strong field-aligned current (FAC) system is located between Saturn and the inner edge of its D-ring, with strength comparable to the high-latitude auroral FACs.

The internal magnetic field of a planet provides a window into its interior [1]. Saturn's magnetic field was first measured by the Pioneer 11 magnetometers [2, 3] and shortly afterward by Voyagers 1 and 2 [4, 5]. Those observations [2–7] revealed that Saturn's magnetic axis is tilted with respect to the planetary spin axis by less than 1°. The axisymmetric part of Saturn's internal field is more complex than a dipole (commonly denoted with the Gauss coefficient g_1^0), with the axisymmetric quadrupole and octupole magnetic moment contributions (coefficients g_2^0 and g_3^0, respectively) amounting to ~10% of the dipole moment contribution when evaluated at the 1-bar surface of the planet. Three decades later, Cassini magnetometer (MAG) measurements [8–12] made before the Grand Finale phase refined Saturn's dipole tilt to <0.06° [11]. Secular variation is at least an order of magnitude slower than at Earth [11], and magnetic flux is expelled from the equatorial region toward mid-to-high latitude [12]. Magnetic moments beyond spherical harmonic degree 3, with Gauss coefficients on the order of 100 nT, were sug-

gested by the MAG measurements during Cassini Saturn Orbit Insertion (SOI) in 2004 [12] but were not well resolved because of the limited spatial coverage of SOI, with periapsis (closest approach) at 1.33 R_S (where 1 R_S = 60,268 km is the 1-bar equatorial radius of Saturn) and latitudinal coverage within ±20° [12].

Saturn's main magnetic field is believed to be generated by rotating convective motion in the metallic hydrogen [13, 14] layer of Saturn through the magnetohydrodynamic (MHD) dynamo process [1, 15–17]. The highly symmetric nature of Saturn's internal magnetic field with respect to the spin axis defies an accurate determination of the rotation rate of Saturn's deep interior and challenges our understanding of the MHD dynamo process because Cowling's theorem [18] precludes the maintenance of a perfectly axisymmetric magnetic field through an active MHD dynamo. The possibility that we are observing Saturn's magnetic field at a time with no active dynamo action can be ruled out given the sizable quadrupole and octupole magnetic moments because these would decay with time at a much faster rate than that of the dipole [[19], chapter 2]. In addition to the main dynamo action in the metallic hydrogen layer, deep zonal flow (differential rotation) and small-scale convective motion in the semiconducting region of Saturn could lead to a secondary dynamo action [20].

We analyzed magnetometer measurements made during the last phase of the Cassini mission. The Cassini Grand Finale phase began on 26 April 2017, with the spacecraft diving through the gap between Saturn's atmosphere and the inner edge of its D-ring 22 times, making a final dive into the atmosphere on 15 September 2017.

Periapses were ~2550 ± 1290 km above the atmosphere's 1-bar level, equivalent to ~1.04 ± 0.02 R_S from the center of Saturn (fig. S1). The low periapses, high inclination (~62°), and proximity to the noon-midnight meridian in local time provided an opportunity to investigate Saturn's internal magnetic field and the magnetic environment between Saturn and its rings (figs. S1 and S2). The fluxgate magnetometer [8] made continuous (32 measurements per second) three-component magnetic field measurements until the final second of spacecraft transmission. For analysis, we averaged the data to 1-s resolution. During ~±50 min around periapsis, the instrument used its highest dynamic range mode, which can measure magnetic field components up to 44,000 nT, with digitization noise of 5.4 nT [8]. This mode had not been used since an Earth flyby in August 1999 [21] and therefore required recalibration during the Grand Finale orbits. Four spacecraft roll campaigns and numerous spacecraft turns enabled this recalibration. The spacecraft attitude needed to be known to a high accuracy of 0.25 millirads, but intermittent suspension of the spacecraft star identification system during phases of instrument targeting resulted in gaps in the required information. These gaps were filled by reconstructing spacecraft attitude using information from the onboard gyroscopes. Attitude reconstruction is currently available for nine of the first 10 orbits, so we restricted our analysis to those orbits. Individual orbits of Cassini are commonly referred to with a rev number; these nine orbits are rev 271 to 280, excluding rev 277.

External magnetic field from sources in the magnetosphere and ionosphere of Saturn

In deriving the internal planetary magnetic field from in situ MAG measurements, it is necessary to separate the internal planetary field from exterior sources of field (currents in the magnetosphere and ionosphere). In the saturnian system, the simplest external source to correct is the magnetodisk current, which flows in the equatorial region far out in the magnetosphere [22–26]. Like the terrestrial ring current, the magnetodisk current produces a fairly uniform depression in the field strength close to the planet and is normally treated as axisymmetric (fig. S3). At Saturn, the magnetodisk current contributes ~15 nT of magnetic field parallel to the spin axis of Saturn (B_Z) in the inner magnetosphere [22–26]. Magnetopause and magnetotail currents, which determine the properties of the outer magnetosphere, contribute ~2 nT B_Z field in the inner magnetosphere of Saturn [22, 26]. More complex are the ubiquitous planetary period oscillations (PPOs), global 10.6- to 10.8-hour oscillations, at close to the expected planetary rotation period that are present in all Saturn magnetospheric parameters [reviewed in [27, 28]]. The PPOs arise from a rotating electrical current systems with sources and sinks in Saturn's ionosphere and magnetosphere and closure currents that flow along magnetic field lines (commonly referred to as FACs)

[1]Physics Department, The Blackett Laboratory, Imperial College London, London, SW7 2AZ, UK. [2]Department of Earth and Planetary Sciences, Harvard University, 20 Oxford Street, Cambridge, MA 02138, USA. [3]Division of Geological and Planetary Sciences, California Institute of Technology, 1200 E California Boulevard, Pasadena, CA 91125, USA. [4]Department of Earth, Planetary, and Space Sciences, University of California, Los Angeles, Los Angeles, CA 90025, USA. [5]Department of Physics and Astronomy, University of Leicester, Leicester LE1 7RH, UK. [6]Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA. [7]Department of Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor, MI 40109, USA.
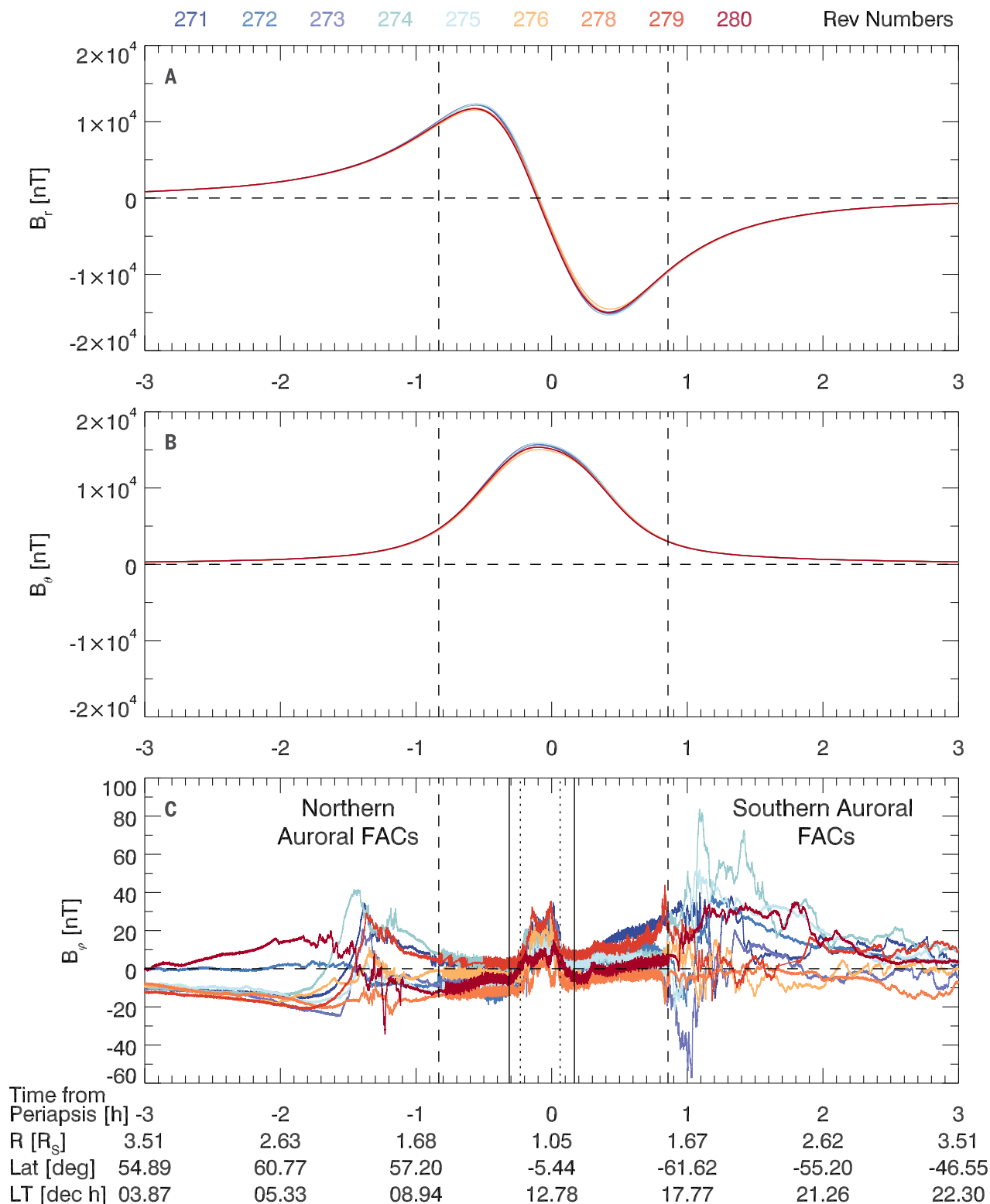*Corresponding author. Email: m.dougherty@imperial.ac.uk

Fig. 1. Vector magnetic field measurements from nine Cassini Grand Finale orbits. (**A** to **C**) The IAU System III Saturn-centered spherical polar coordinates are adopted here. The peak measured magnetic field strength is ~18,000 nT, whereas the azimuthal component of the field (plotted on a different scale) is ~1/1000 of the total field outside the high-latitude auroral FACs region, which are labeled. Vertical dashed lines indicate the MAG range 3 ($|B| > 10,000$ nT) time period. (C) The vertical dotted lines mark the inner edge of the D-ring mapped magnetically, and the solid vertical solid lines mark the outer edge of the D-ring mapped magnetically, as described in the text.

and map to the auroral regions. Normally, two periods are present at any one time, with one dominant in the northern hemisphere and the other in the southern (29–33), which during the interval studied here were 10.79 and 10.68 hours, respectively (33). The rotation periods, amplitudes, and relative amplitudes of the two PPO systems vary slowly over Saturn's seasons (34–36). These seasonal changes indicate that the source of the periods is most likely to be of atmospheric origin rather than deep within the planet (37, 38). Moreover, the PPO periods are a few percent longer compared with any likely rotation period of the deep interior of the planet (39–41). There is also a nonoscillatory system of strong FACs in the auroral region with current closure through the planetary ionosphere, and there are distributed FACs at other latitudes (42–44). This nonoscillatory current system is associated with magnetospheric plasma that rotates more slowly than the planet (42). During the Grand Finale phase, the Cassini spacecraft was close enough to the planet during periapse to directly detect currents in the ionosphere. We describe results on these external current systems first because they must be considered during analysis of the internal field.

Shown in Fig. 1 is the vector magnetic field measurements in International Astronomical Union (IAU) Saturn System III right-hand spherical polar coordinates, with r, θ, and φ denoting radial, meridional, and azimuthal directions, respectively (45) from the nine Cassini Grand Finale orbits for which accurate spacecraft attitude information is available. This coordinate system has its origin at the center of mass of Saturn, and the spin axis of Saturn is the polar axis (zenith reference). The peak magnetic field strength encountered during these Grand Finale orbits is ~18,000 nT, which is well within the range of the fluxgate magnetometer (8). The peak field was not encountered during periapsis but at mid-latitude in the southern hemisphere. The dominant radial and meridional magnetic field components, B_r and B_θ, which exhibit similar behavior on each orbit, are shown in Fig. 1, A and B. This repeatability, combined

Fig. 2. Cassini's Grand Finale trajectory shown in the meridional plane. The trajectory of Cassini Grand Finale orbit rev 275 (black/blue trace with arrows) in the ρ-Z cylindrical coordinate system—where ρ is the perpendicular distance from the spin axis, and Z is distance from Saturn's planetary equator—is overlain on the average positions of Saturn's auroral FACs region (gray region) and the newly discovered low-latitude FACs (red region). Thin gray traces with arrows are magnetic field lines. The interval along the trajectory during which the measured magnetic field was >10,000 nT is highlighted in blue. Small circles are at 3-hour intervals, and the beginning of day 142 of year 2017 is shown.
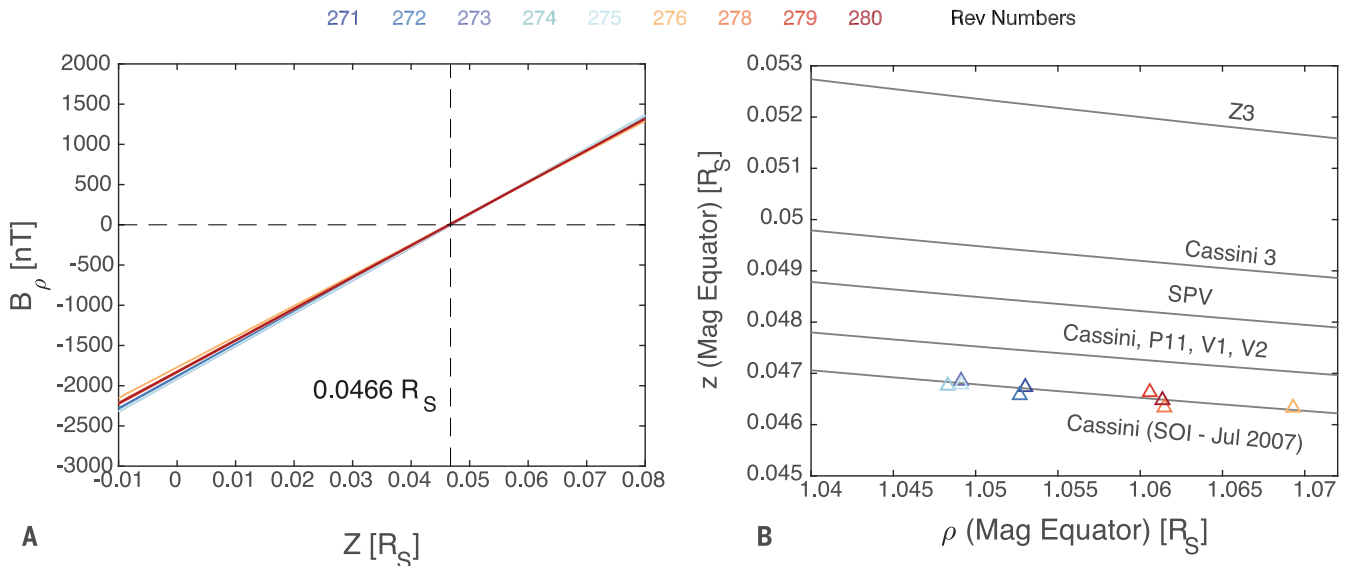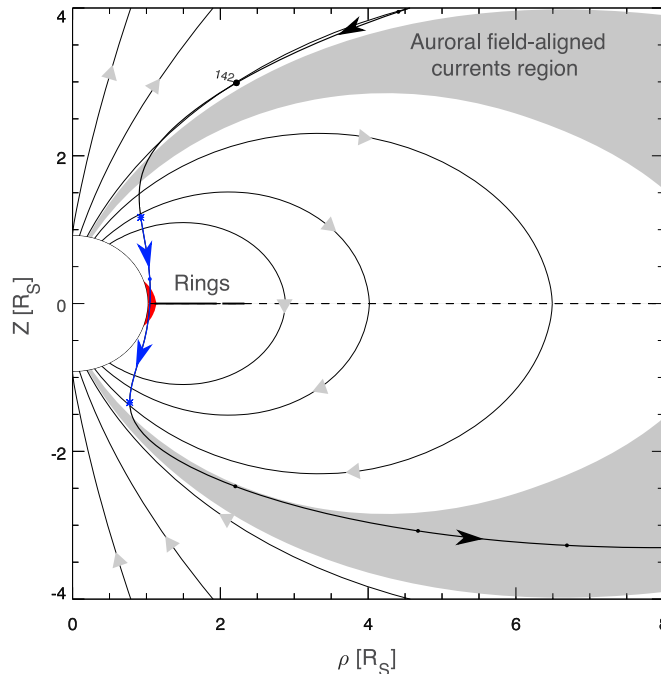

Fig. 3. Saturn's magnetic equator position as directly measured along the Cassini Grand Finale orbits. (**A**) The measured cylindrical radial component of the magnetic field versus distance from Saturn's planetary equator from nine Cassini Grand Finale orbits. (**B**) The distribution of the magnetic equator northward displacements (triangles) versus cylindrical radial distance compared with predictions from existing degree-3 internal field models (lines) (4, 7, 10, 11). The Z3 model is from (4), the Cassini 3 model is from (11), the SPV model is from (7), the Cassini (SOI, June 2007) model, and the Cassini, P11, V1, V2 model are both from (10). The Cassini (SOI, June 2007) model were derived from Cassini SOI data to July 2007 data, and the Cassini, P11, V1, V2 model is derived from Cassini SOI to July 2007 data combined with Pioneer 11 and Voyager 1 and 2 data (10).

with the fixed local time of the orbits and the slight differences in orbital period, demonstrate that the planetary magnetic field is close to axisymmetric.

Shown in Fig. 1C is the azimuthal component of the field, B_φ, which is almost three orders of magnitude smaller than the radial and meridional components outside the auroral FACs region. If it originates from the interior of the planet, B_φ would be expected to decay with radial distance and also be part of a nonaxisymmetric field pattern (Eq. 5). However, much of the observed B_φ signal seems to be of external rather than internal origin. The magnetic perturbations from the northern and southern auroral currents through which the spacecraft flew are marked in Figs. 1 and 2. In Fig. 2, the cylindrical coordinate system is adopted, where ρ is the perpendicular distance to the spin axis of Saturn and Z is the distance from the planetary equator of Saturn, which is defined by its center of mass. These current sheets have been crossed many times during the Cassini mission, but previous crossings occurred much further from the planet (*43*, *44*, *46*). The sharp field gradients observed are due to local FACs, and the short scale of the variation in the field indicates local current sources rather than relatively distant currents in the planet's interior. The scatter in the auroral FACs from orbit to orbit is partly due to the variability of the aurora over time and partly due to the large contribution from the rotating PPO signals, which are encountered at different phases on each orbit.

The narrow central peak in B_φ ranging between 5 and 30 nT observed along these Grand Finale orbits was unexpected. It can be seen from Fig. 1C that the large change in field slope takes place where the spacecraft crosses the magnetic field lines that map to the inner edge of the D-ring. We adopted 1.11 R_S from the center of Saturn in the equatorial plane as the D-ring inner edge. If the azimuthal field gradient is due to a local FAC, the current within the D-ring is interhemispheric and always flows from the northern to the southern hemisphere along these nine Grand Finale orbits. There must be return currents either at lower altitude than the spacecraft penetrates in the ionosphere or at other local times. Assuming approximate axisymmetry (for azimuthal spatial scales much larger than latitudinal scales), Ampère's law combined with current continuity leads to a total interhemispheric current flow in the range of 0.25 million to 1.5 million ampere per radian of azimuth. Thus, this low-latitude FAC is comparable in strength with the currents observed on field lines connected to the high-latitude auroral region (*43*, *44*, *46*). It thus may be part of a global magnetospheric–ionospheric coupling system. Alternatively, the system might be purely of ionospheric origin, in which case it could be due to skewed flow between northern and southern ends of the field line (*47*). All the orbit periapses are near local noon, so the repeatability from periapsis to periapsis does not rule out a dependency on local time.
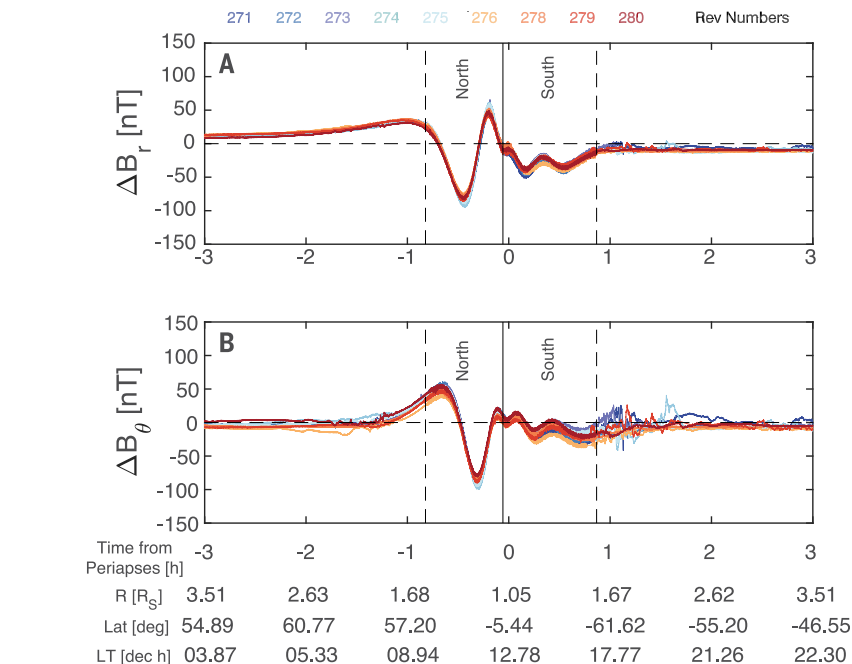

Fig. 4. Saturn's magnetic field beyond spherical harmonic degree 3 as directly measured during the Grand Finale orbits. (**A** and **B**) Residuals of the radial and meridional magnetic fields from the unregularized degree-3 internal field model (parameters are provided in table S1). It can be seen that the residuals are on the order of 100 nT, which is 10 times larger than the magnetodisk field.

There is a clear difference between the signatures of northern and southern auroral FACs (Fig. 1C), and again, magnetic mapping reveals the source. The locations of the auroral current sheets are well known magnetically from previous Cassini measurements (Fig. 2) (*43*, *44*, *46*). During Cassini's inbound trajectory in the north, the spacecraft is located on the field lines just poleward of the main FAC region, as shown by the decreasing B_φ, which is indicative of the distributed downward current on polar field lines, as previously shown (*43*, *44*, *46*). The less structured B_φ during the outbound southern trajectories (Fig. 1C) indicates that the southern hemisphere auroral FACs are not being fully crossed by the spacecraft, and instead we observed the center of FAC activity moving dynamically back and forth over the spacecraft.

Saturn's internal magnetic field

Next we analyzed the radial and meridional components of the measured magnetic field, which are dominated by the axisymmetric internal magnetic field. The highly inclined nature of the Grand Finale orbits allows direct determination close to the planet of Saturn's magnetic equator, which is defined as where the cylindrical radial component of the magnetic field (B_ρ) vanishes. The measured B_ρ as a function of the spacecraft's distance from the planetary equator of Saturn is shown in Fig. 3A. The measured value along these orbits shows that Saturn's magnetic equator is 0.0466 ± 0.0002 (1 SD) R_S northward of the planetary equator, measured at distance ~1.05 R_S from the spin axis of Saturn.

This indicates a small yet non-negligible north-south asymmetry in Saturn's magnetic field. Shown in Fig. 3B are the distribution of the directly measured magnetic equator position of Saturn in the ρ-Z plane and predictions from various preexisting degree-3 models (*4*, *7*, *10*, *11*). Several degree-3 internal field models have been derived in (*10*) from different combinations of datasets; the two shown in Fig. 3 were (i) derived from data between Cassini SOI to July 2007 and (ii) data from Cassini SOI to July 2007 combined with Pioneer 11 and Voyager 1 and 2 data. A nominal magnetodisk field (table S2) was added to the internal field when predicting the magnetic equator positions from existing models. A systematic decrease of the magnetic equator position with distance from the spin axis is evident. This is a consequence of the axisymmetric part of the magnetic field, as illustrated by the predictions from existing axisymmetric models. In addition to this systematic trend, variations of measured magnetic equator positions at similar ρ were also observed. These additional variations of the magnetic equator positions serve as a direct bound on possible nonaxisymmetry of Saturn's internal magnetic field because any internal nonaxisymmetry would cause longitudinal variations of the magnetic equator positions. The measured peak-to-peak variation at similar cylindrical radial distances, ~0.00030 R_S (18 km), translates to a dipole tilt of 0.0095°, a guide value for possible internal nonaxisymmetry. This value is about one order of magnitude smaller than the upper limit on dipole tilt placed before the Grand Finale (*10*, *11*, *48*). The

corresponding field components at spacecraft altitude are only ~3 nT (in a background field of 18,000 nT), which is too small to be discerned from external fields in the measured B_φ. Such an extreme axisymmetric planetary magnetic field is difficult to explain with a dynamo model ([18], [49–52]).

Although the internal magnetic field of Saturn is highly axisymmetric (structureless in the longitudinal direction), it exhibits latitudinal structures across many different length scales. This is revealed by our retrieval of the axisymmetric Gauss coefficients of Saturn's internal magnetic field from these nine Grand Finale orbits. In this analysis, the magnetodisk field was explicitly included, adopting an analytical formula ([25]) with initial parameters taken from ([22]). The axisymmetric internal Gauss coefficients and the magnetodisk current field were then retrieved via an iterative process. To determine the internal Gauss coefficients, only measurements from range 3 of the fluxgate magnetometer were adopted ($|B| > 10,000$ nT, $r < 1.58\ R_S$), and measurements from all nine orbits were treated as a single dataset. These measurements were outside the auroral FACs region, as shown in Figs. 1 and 2. For retrieval of the magnetodisk current field, measurements from both range 3 and range 2 of the fluxgate magnetometer were used ($|B| > 400$ nT, $r < 4.50\ R_S$), and each orbit was treated separately (table S2). Along the trajectory of the close-in portion of the Cassini Grand Finale orbits, the contributions from the magnetopause and magnetotail currents cannot be practically separated

from the quasi-uniform magnetodisk current contribution and thus are subsumed into our magnetodisk current model. The time-varying external (magnetodisk + magnetopause + magnetotail) B_Z field could create a time-varying component of the internal dipole g_1^0 through electromagnetic induction. The maximum magnetodisk field variation we have observed along the nine orbits is ~10 nT, which for an induction depth of 0.87 R_S (Methods and fig. S4) would induce a time variation in the internal dipole coefficient of ~3.3 nT. This is less than the digitization level of the highest range of the fluxgate magnetometer onboard Cassini (5.4 nT); thus, the detection of which is not straightforward.

We first investigated the minimum parameter set (in terms of internal Gauss coefficients) needed to adequately describe the observations, while applying no regularization to the parameters (Methods) ([53], [54]). This experiment revealed that axisymmetric internal Gauss coefficients up to degree 9 are required to bring the root-mean-square (RMS) residual in range 3 below 10 nT (fig. S5). The residual (B_r, B_θ) after removal of the unregularized degree 3 model (coefficients provided in table S1) are on the order of 100 nT but have larger amplitude and larger spatial scales in the northern hemisphere as compared with those in the southern hemisphere (Fig. 4). Shown in Fig. 5 is the residual (B_r, B_θ) as a function of latitude after removal of the unregularized degree 6 model and the best-fitting magnetodisk model for each orbit (coefficients provided in tables S1 and S2), which exhibits consistent small-scale features with am-

plitude ~25 nT and typical spatial scale ~25° in latitude. The highly consistent nature (from orbit to orbit) of these small-scale magnetic perturbations in (B_r, B_θ), in contrast to the orbit-to-orbit varying B_φ signature, suggests a source below the highly time-variable currents in the ionosphere of Saturn.

A new internal magnetic field model for Saturn, which we refer to as the Cassini 11 model, is presented in Table 1 and Fig. 6, with 11 denoting the highest spherical harmonic degree Gauss coefficient above the uncertainties (Methods). Because there was incomplete latitudinal coverage during the Grand Finale orbits, which can lead to nonuniqueness in the solution, a regularized inversion technique ([53], [54]) has been used to construct the Cassini 11 model. With the regularized inversion technique ([53], [54]), we explicitly sought internal field solutions that not only match the observations but also contain minimum magnetic power beyond degree 3, when evaluated at a reference radius (Methods). Presented in Table 1 are the axisymmetric Gauss coefficients, five times the formal uncertainty (5σ) (definition is provided in Methods), and the RMS residual of the Cassini 11 model. The resulting Mauersberger-Lowes (M-L) magnetic power spectrum ([55–57]) of the Cassini 11 model evaluated at the surface of Saturn is shown in Fig. 6. In the Cassini 11 model, the axisymmetric Gauss coefficients between degree 4 and degree 11 are on the order of 10 to 100 nT, whereas those above degree 11 are below the 5σ uncertainty.

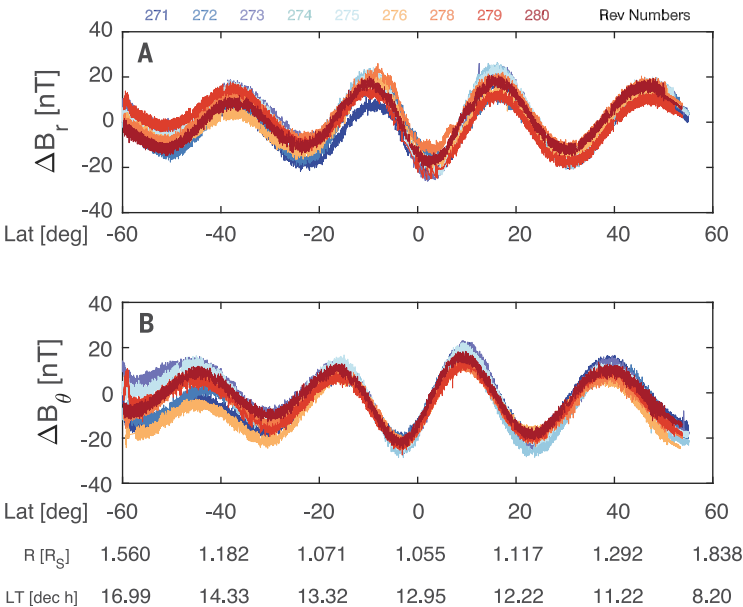We have also investigated whether it is possible to obtain a simpler description of Saturn's

Fig. 5. Saturn's magnetic field beyond spherical harmonic degree 6 as directly measured during the Grand Finale. (**A** and **B**) Residuals of the radial and meridional magnetic fields from the unregularized degree-6 internal field model (parameters are provided in table S1). The field corresponding to the magnetodisk current has also been removed (parameters of the magnetodisk current model are provided in table S2). Latitudinally banded magnetic structures on the order of 25 nT are evident. Orbit-to-orbit deviations in B_θ between −60° and −40° latitudes are mostly due to the influences of southern hemisphere high-latitude FACs.

Table 1. Gauss coefficients of a new model for Saturn's internal magnetic field, which we refer to as the Cassini 11 model, constructed from nine orbits of Cassini Grand Finale MAG data with regularized inversion. The reported uncertainty is five times the formal uncertainties associated with the chosen regularization (Methods).

Gauss coefficient	Value (nT)	Uncertainty (nT)
g_1^0	21140.2	1.0
g_2^0	1581.1	1.2
g_3^0	2260.1	3.2
g_4^0	91.1	4.2
g_5^0	12.6	7.1
g_6^0	17.2	8.2
g_7^0	59.6	8.1
g_8^0	−10.5	8.7
g_9^0	−12.9	6.3
g_{10}^0	15.0	7.0
g_{11}^0	18.2	7.1
g_{12}^0	0.3	7.7
RMS residual	6.2	

internal magnetic field (in terms of Gauss co-efficients) in a shifted coordinate system. We shifted the coordinate system northward from Saturn's center of mass along the direction of the spin axis, in the range of 0.03 to 0.05 R_S denoted by z_S, and solved for the internal Gauss coefficients. We found that no simpler description of the field in terms of Gauss coefficients can be obtained in the Z-shifted coordinates because of the complex nature of Saturn's internal magnetic field in the latitudinal direction. Only the quadrupole moment g_2^0 can be reduced to zero in the Z-shifted coordinates with $z_S \sim 0.0337$ R_S; all other even-degree moments (such as g_4^0, g_6^0, and g_8^0) remain nonzero (fig. S6). In addition, g_4^0 becomes more than an order of magnitude larger in these shifted coordinates compared with that in the Saturn-centered coordinates. The RMS residual remains unchanged as we shifted the coordinate system. We therefore did not consider models in shifted coordinates any further.

Implications for Saturn's dynamo and interior

Mathematically, downward continuation of the magnetic field measured outside of a planet toward its interior is strictly valid through regions with no electrical currents **J**. From the perspective of dynamo action, the downward continuation of the magnetic field is possible through regions with no substantial dynamo action. For Saturn, because of the smooth increase of electrical conductivity as a function of depth ([13], [14], [58]), the vigor of dynamo action is expected to rise smoothly with depth ([20]) and can be quantified by the magnetic Reynolds number (R_m), which measures the ratio of magnetic field production/modification by flow in an electrically conducting fluid against ohmic diffusion (Methods) ([59]). Assuming 1 cm s^{-1} flow and an electrical conductivity model from ([58], [60]), the local R_m reaches 10 at 0.77 R_S, 50 at 0.725 R_S, and 100 at 0.70 R_S (fig. S8). For dynamo action in the shallow semiconducting layer, $R_m = 20$ is likely to be sufficient given the existence of the main magnetic field originating from the deep dynamo. Thus, the top of the shallow dynamo is likely to fall between 0.77 and 0.725 R_S. We chose 0.75 R_S as an approximation for the surface of the shallow dynamo; the properties of the downward continued magnetic field are broadly similar in this range of depths.

The small-scale B_r beyond degree 3 within ±60° latitude at 0.75 R_S based on the Cassini 11 model is shown in Fig. 7. The small-scale magnetic structures measured along the spacecraft trajectory map to latitudinally banded magnetic structures with amplitude ~5000 nT and typical spatial scale ~15° in latitude at 0.75 R_S, which amount to ~5 to 10% of the local background field (fig. S7). The inferred small-scale magnetic field pattern is broadly similar for either a spherical surface with radius 0.75 R_S or a dynamically flattened elliptical surface with equatorial radius 0.75 R_S and polar radius 0.6998 R_S (fig. S9). The magnetic power in the low-degree

Fig. 6. Magnetic power spectrum of the Cassini 11 model. This new internal field model for Saturn is constructed from nine Cassini Grand Finale orbits with regularized inversion. Central values and five times the formal uncertainties derived from the regularized inversion are shown (Methods). It can be seen that the high-degree moments between degree 4 and degree 11 are on the order of 10 to 100 nT, whereas those above degree 11 are below the derived uncertainty.
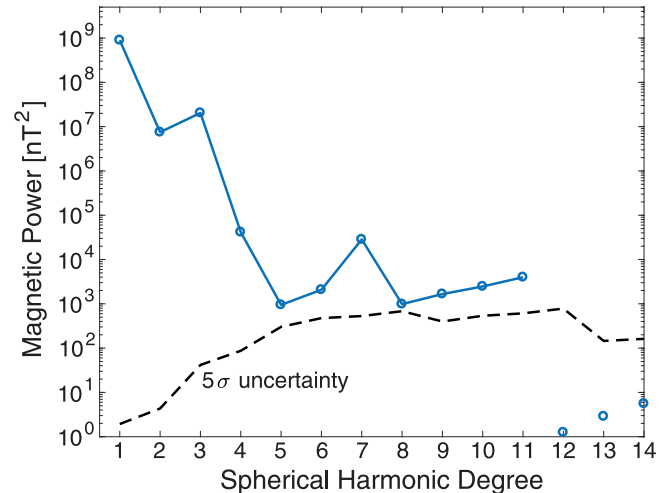
Fig. 7. Small-scale axisymmetric magnetic field of Saturn at 0.75 R_S. The values of ΔB_r were computed by using the central values of the degree 4 to degree 11 Gauss coefficients (Table 1). These small-scale magnetic perturbations at 0.75 R_S are on the order of 5000 nT, which amount to ~5 to 10% of the local background field (fig. S6). The plotted latitudinal range corresponds to the region of Cassini MAG range 3 measurements along the Grand Finale orbits.
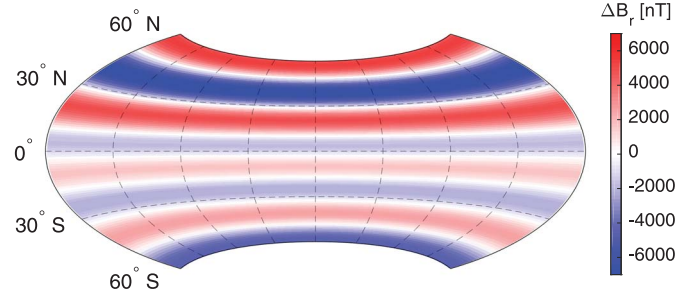
(degrees 1 to 3) part of the M-L spectrum still greatly exceeds that in the higher degree part when evaluated at 0.75 R_S; the dipole power is one order of magnitude larger than the octupole power, whereas the octupole power is almost two orders of magnitude larger than those of degrees 4 to 11. The orders-of-magnitude difference between the low-degree power and high-degree power at 0.75 R_S suggests that the latitudinally banded magnetic perturbations could originate from a secondary dynamo action in the semiconducting region of Saturn ([20]). This secondary dynamo action likely only produces perturbations on top of the primary deep dynamo, as suggested by a kinematic mean-field dynamo model ([20]). Fully three-dimensional (3D) self-consistent giant planet dynamo models with deep differential rotation (zonal flows) also produced latitudinally banded magnetic fields ([61], [62]).

The dynamo α-effect, which describes the generation of a poloidal magnetic field from a toroidal magnetic field [poloidal and toroidal magnetic field are defined in ([63])] by rotating convective motion, is likely to be of the traditional mean-field type in the shallow dynamo ([16], [17], [20]). The validity of the traditional α-effect

has been extensively discussed in ([20]). In this scenario, no substantial low-degree nonaxisymmetric magnetic field is expected from the shallow dynamo.

From an empirical perspective, some but not all of the high-degree magnetic moments could still have a deep interior origin. One extreme interpretation of the Cassini 11 model is to regard the magnetic moments up to degree 9 as entirely being of deep interior origin because the magnetic power of degrees 7 and 9 is within a factor of three of that of degree 3 when evaluated at 0.50 R_S. When discussing internal magnetic fields well below 0.75 R_S, we have (i) treated the shallow dynamo as a small perturbation and (ii) implicitly assumed stable stratification between the surface of the deep dynamo and the shallow dynamo. Under this assumption, there exists no local dynamo α-effect between the surface of the deep dynamo and the shallow dynamo. Although the ω-effect, which describes the generation of the toroidal magnetic field from differential rotation acting on the poloidal magnetic field, likely operates in the stably stratified layer, the ω-effect will not affect the poloidal component of the magnetic field. Downward continuation of the magnetic

field well below 0.75 R_S can only be justified if there are no effective electrical currents generating the poloidal magnetic field in the layer between the surface of the deep dynamo and the shallow dynamo.

The radial magnetic field is shown in fig. S10 as a function of latitude at 0.50 R_S computed from the Gauss coefficients of the Cassini 11 model up to degree 9. It can be seen that the radial component of Saturn's magnetic field at 0.50 R_S is strongly concentrated toward high latitude, leaving minimum magnetic flux at latitudes below ±45°, a weaker version of which is evident from the magnetic moments up to degree 3 (fig. S10). The possibility of weak magnetic field at the poles of Saturn near 0.50 R_S is suggested in fig. S10. In this alternative interpretation, magnetic moments beyond degree 9 would be of shallower origin. This highlights the ambiguity in the separation of magnetic moments of shallow origin from those of deep origin, which cannot be resolved with MAG data alone.

The main dynamo of Saturn does appear to be deeply buried, originating from a depth deeper than the expected full metallization of hydrogen at ~0.64 R_S (13). The surface strength of Saturn's internal magnetic field is more than one order of magnitude weaker than that of Jupiter (64–67). From dynamo scaling analysis, Saturn's magnetic field appears to be weaker than expected from both force-balance considerations and energy-balance considerations if the outer boundary of the deep dynamo is much shallower than 0.40 R_S (60). The most widely accepted theoretical explanation for Saturn's highly axisymmetric magnetic field (49, 50) suggests that differential rotation in a stably stratified and electrically conducting layer above the deep dynamo region electromagnetically filters out nonaxisymmetric magnetic moments. Although 3D numerical dynamo simulations (51, 52) that implement stable stratification and differential rotation have only produced an average dipole tilt of 0.6°, about two orders of magnitude larger than the newly derived guide value 0.0095°, the R_m in these 3D numerical dynamo simulations (51, 52) are only in the range of 100 to 500, which is orders of magnitude smaller than possible values (~30,000) inside the metallic hydrogen layer of Saturn (68). At present, we regard the extreme axisymmetry of Saturn's magnetic field as a continuing enigma of the inner workings of Saturn.

Methods
Un-regularized and regularized inversion of Saturn's internal magnetic field

The internal magnetic field of Saturn outside the dynamo region is described using Gauss coefficients (g_n^m, h_n^m) where n and m are the spherical harmonic degree and order, respectively:

$$V^{\text{Int}} = \sum_{n=1}\sum_{m=0}^{l} R_p \left(\frac{R_p}{r}\right)^{n+1}$$
$$[g_n^m \cos m\varphi + h_n^m \sin m\varphi] P_n^m(\cos\theta) \quad (1)$$

$$(B_r^{\text{Int}}, B_\theta^{\text{Int}}, B_\varphi^{\text{Int}}) = -\nabla V^{\text{Int}} \quad (2)$$

$$B_r^{\text{Int}} = \sum_{n=1}\sum_{m=0}^{n} (n+1)\left(\frac{R_p}{r}\right)^{n+2}$$
$$[g_n^m \cos m\varphi + h_n^m \sin m\varphi] P_n^m(\cos\theta) \quad (3)$$

$$B_\theta^{\text{Int}} = -\sum_{n=1}\sum_{m=0}^{n}\left(\frac{R_p}{r}\right)^{n+2}$$
$$[g_n^m \cos m\varphi + h_n^m \sin m\varphi] \frac{dP_n^m(\cos\theta)}{d\theta} \quad (4)$$

$$B_\varphi^{\text{Int}} = \sum_{n=1}\sum_{m=0}^{n}\left(\frac{R_p}{r}\right)^{n+2}\frac{m}{\sin\theta}$$
$$[g_n^m \sin m\varphi - h_n^m \cos m\varphi] P_n^m(\cos\theta) \quad (5)$$

It can be seen from Eq. 5 that the azimuthal magnetic field of internal origin, B_φ^{Int}, must be part of a non-axisymmetric ($m \neq 0$) field pattern.

The forward model can be represented as

$$\boldsymbol{data} = G \, \boldsymbol{model} \quad (6)$$

For this particular problem, **data** represents the magnetic field measurements (B_r, B_θ, B_φ), **model** represents the Gauss coefficients (g_n^m, h_n^m), and G represents the matrix expression of Eqs. 3 to 5.

In un-regularized inversion, one seeks to minimize the difference between the data and the model field only

$$|\boldsymbol{data} - G\,\boldsymbol{model}|^2 \quad (7)$$

whereas in regularized inversion, additional constraints are placed on the model parameters. Instead, one seeks to minimize

$$|\boldsymbol{data} - G\,\boldsymbol{model}|^2 + \gamma^2 |L\,\boldsymbol{model}|^2 \quad (8)$$

in which L represents a particular form of regularization and γ is a tunable damping parameter adjusting the relative importance of model constraints.

We seek to minimize the power in the surface integral of the radial component of the magnetic field beyond spherical harmonic degree 3

$$\int_{r=r_{\text{ref}}} B_r^2(n>3)\,d\Omega = \sum_{n=4}\frac{(n+1)^2}{2n+1}\left(\frac{R_p}{r_{\text{ref}}}\right)^{2n+4}$$
$$[(g_n^m)^2 + (h_n^m)^2] \quad (9)$$

at a reference radius r_{ref}. Thus,

$$L = \frac{n+1}{\sqrt{2n+1}}\left(\frac{R_p}{r_{\text{ref}}}\right)^{n+2} \quad (10)$$

for $n > 3$ and $L = 0$ for $n \leq 3$.

After experimenting with different r_{ref} and γ, we choose $r_{\text{ref}} = 1.0$ and $\gamma = 0.6$ which yield smooth magnetic field perturbations when viewed between 1.0 R_S and 0.75 R_S.

In this framework, the model solution can be computed via

$$\boldsymbol{model} = (G^T G + \gamma^2 L^T L)^{-1} G^T \boldsymbol{data} = G^N \boldsymbol{data} \quad (11)$$

in which superscript T represents matrix transpose.

The model covariance matrix can be computed from

$$mod_{\text{cov}} = G^N \, data_{\text{cov}} \, (G^N)^T \quad (12)$$

where $data_{\text{cov}}$ is the data covariance matrix, taken to be $\sigma_{\text{data}}^2 \boldsymbol{I}$, in which σ_{data} is the RMS of the data model misfit and $\boldsymbol{I}$ is the identity matrix (54). The formal uncertainty of the model solution is the square root of the diagonal term in the model covariance matrix.

Electromagnetic induction inside Saturn

Due to the rapidly radial varying nature of electrical conductivity inside Saturn, the three factors controlling the electromagnetic (EM) induction at Saturn are: the frequency of the inducing field ω_{ind}, the electrical conductivity $\sigma(r)$ and the electrical conductivity scale height $H_\sigma = \sigma/|d\sigma/dr|$. The EM induction inside Saturn would occur at a depth where the frequency dependent local EM skin depth $d(r, \omega_{\text{ind}}) = \sqrt{2/\omega_{\text{ind}}\mu_0\sigma(r)}$ becomes similar to or less than the local electrical conductivity scaleheight $H_\sigma(r)$, where μ_0 is the magnetic permeability.

Figure S4 shows the frequency dependent EM skin depth compared to the electrical conductivity scaleheight inside Saturn, adopting the electrical conductivity profile (50). Two frequencies have been chosen, the rotational frequency of Saturn $\omega_r = 2\pi/10.5$ hours ~1.7 × 10⁻⁴ rad s⁻¹ and the orbital period of Cassini Grand Finale orbits $\omega_o = 2\pi/7$ Earth days ~1.0 × 10⁻⁵ rad s⁻¹. It can be seen from fig. S4 that EM induction inside Saturn would occur at depth 0.87 R_S for ω_{ind} equal to the rotational frequency of Saturn and at depth 0.86 R_S for ω_{ind} equal to the orbital frequency of Cassini Grand Finale orbits. Thus, the induction depth inside Saturn would likely be 0.87 R_S or less.

A 10 nT inducing B_Z field would create a time-varying internal dipole field with equatorial strength of 5 nT and polar strength of 10 nT when evaluated at this depth, which corresponds to a time varying dipole coefficient g_1^0 of 3.3 nT when defined with respect to the equatorial radius of Saturn's 1 bar surface.

REFERENCES AND NOTES

1. D. J. Stevenson, Planetary magnetic fields. *Earth Planet. Sci. Lett.* **208**, 1–11 (2003). doi: 10.1016/S0012-821X(02)01126-3
2. E. J. Smith et al., Saturn's magnetic field and magnetosphere. *Science* **207**, 407–410 (1980). doi: 10.1126/science.207.4429.407; pmid: 17833549
3. M. H. Acuña, N. F. Ness, The magnetic field of saturn: Pioneer 11 observations. *Science* **207**, 444–446 (1980). doi: 10.1126/science.207.4429.444; pmid: 17833558
4. J. E. P. Connerney, N. F. Ness, M. H. Acuña, Zonal harmonic model of Saturn's magnetic field from Voyager 1 and 2 observations. *Nature* **298**, 44–46 (1982). doi: 10.1038/298044a0
5. J. E. P. Connerney, M. H. Acuña, N. F. Ness, The Z3 model of Saturn's magnetic field and the Pioneer 11 vector

helium magnetometer observations. *J. Geophys. Res.* **89** (A9), 7541–7544 (1984). doi: 10.1029/JA089iA09p07541

6. L. Davis Jr., E. J. Smith, New models of Saturn's magnetic field using Pioneer 11 Vector Helium Magnetometer data. *J. Geophys. Res.* **91** (A2), 1373–1380 (1986). doi: 10.1029/JA091iA02p01373

7. L. Davis Jr., E. J. Smith, A model of Saturn's magnetic field based on all available data. *J. Geophys. Res.* **95** (A9), 15257–15261 (1990). doi: 10.1029/JA095iA09p15257

8. M. K. Dougherty *et al.*, The Cassini magnetic field investigation. *Space Sci. Rev.* **114**, 331–383 (2004). doi: 10.1007/s11214-004-1432-2

9. M. K. Dougherty *et al.*, Cassini magnetometer observations during Saturn orbit insertion. *Science* **307**, 1266–1270 (2005). doi: 10.1126/science.1106098; pmid: 15731444

10. M. E. Burton, M. K. Dougherty, C. T. Russell, Models of Saturn's internal planetary magnetic field based on Cassini observations. *Planet. Space Sci.* **57**, 1706–1713 (2009). doi: 10.1016/j.pss.2009.04.008

11. H. Cao, C. T. Russell, U. R. Christensen, M. K. Dougherty, M. E. Burton, Saturn's very axisymmetric magnetic field: No detectable secular variation or tilt. *Earth Planet. Sci. Lett.* **304**, 22–28 (2011). doi: 10.1016/j.epsl.2011.02.035

12. H. Cao, C. T. Russell, J. Wicht, U. R. Christensen, M. K. Dougherty, Saturn's high degree magnetic moments: Evidence for a unqiue planetary dynamo. *Icarus* **221**, 388–394 (2012). doi: 10.1016/j.icarus.2012.08.007

13. S. T. Weir, A. C. Mitchell, W. J. Nellis, Metallization of fluid molecular hydrogen at 140 GPa (1.4 Mbar). *Phys. Rev. Lett.* **76**, 1860–1863 (1996). doi: 10.1103/PhysRevLett.76.1860; pmid: 10060539

14. M. Zaghoo, I. F. Silvera, Conductivity and dissociation in liquid metallic hydrogen and implications for planetary interiors. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 11873–11877 (2017). doi: 10.1073/pnas.1707918114; pmid: 29078318

15. E. N. Parker, Hydromagnetic dynamo models. *Astrophys. J.* **122**, 293 (1955). doi: 10.1086/146087

16. F. Krause, K.-H. Rädler, *Mean-Field Magnetohydrodynamics and Dynamo Theory* (Pergamon, 1980).

17. H. K. Moffatt, *Magnetic Field Generation in Electrically Conducting Fluids* (Cambridge Univ. Press, 1978).

18. T. G. Cowling, The magnetic field of sunspots. *Mon. Not. R. Astron. Soc.* **94**, 39–48 (1933). doi: 10.1093/mnras/94.1.39

19. P. Charbonneau, *Solar and Stellar Dynamos: Saas-Fee Advanced Course 39 Swiss Society for Astrophysics and Astronomy* (Springer-Verlag Berlin Heidelberg, 2013).

20. H. Cao, D. J. Stevenson, Zonal flow magnetic field interaction in the semi-conducting region of giant planets. *Icarus* **296**, 59–72 (2017). doi: 10.1016/j.icarus.2017.05.015

21. D. J. Southwood *et al.*, Magnetometer measurements from the Cassini Earth swing-by. *J. Geophys. Res.* **106** (A12), 30109–30128 (2001). doi: 10.1029/2001JA900110

22. E. J. Bunce *et al.*, Cassini observations of the variation of Saturn's ring current parameters with system size. *J. Geophys. Res.* **112** (A10), A10202 (2007). doi: 10.1029/2007JA012275

23. J. E. P. Connerney, M. H. Acuña, N. F. Ness, Saturn's ring current and inner magnetosphere. *Nature* **292**, 724–726 (1981). doi: 10.1038/292724a0

24. J. E. P. Connerney, M. H. Acuña, N. F. Ness, Currents in Saturn's magnetosphere. *J. Geophys. Res.* **88** (A11), 8779–8789 (1983). doi: 10.1029/JA088iA11p08779

25. G. Giampieri, M. Dougherty, Modeling of the ring current in Saturn's magnetosphere. *Ann. Geophys.* **22**, 653–659 (2004). doi: 10.5194/angeo-22-653-2004

26. T. M. Edwards, E. J. Bunce, S. W. H. Cowley, A note on the vector potential of Connerney et al.'s model of the equatorial current sheet in Jupiter's magnetosphere. *Planet. Space Sci.* **49**, 1115–1123 (2001). doi: 10.1016/S0032-0633(00)00164-1

27. J. F. Carbary, D. G. Mitchell, Periodicities in Saturn's magnetosphere. *Rev. Geophys.* **51**, 1–30 (2013). doi: 10.1002/rog.20006

28. D. J. Southwood, S. W. H. Cowley, The origin of Saturn magnetic periodicities: Northern and southern current systems. *J. Geophys. Res.* **119**, 1563–1571 (2014). doi: 10.1002/2013JA019632

29. D. A. Gurnett *et al.*, Discovery of a north-south asymmetry in Saturn's radio rotation period. *Geophys. Res. Lett.* **36**, L16102 (2009). doi: 10.1029/2009GL039621

30. D. J. Andrews *et al.*, Magnetospheric period oscillations at Saturn: Comparison of equatorial and high latitude magnetic field periods with north and south Saturn kilometric radiation periods. *J. Geophys. Res.* **115** (A12), 1363 (2010). doi: 10.1029/2010JA015666

31. D. J. Southwood, Direct evidence of differences in magnetic rotation rate between Saturn's northern and southern polar regions. *J. Geophys. Res.* **116** (A1), 1–11 (2011). doi: 10.1029/2010JA016070

32. G. J. Hunt, G. Provan, S. W. H. Cowley, M. K. Dougherty, D. J. Southwood, Saturn's planetary period oscillations during

33. the closest approach of Cassini's ring-grazing orbits. *Geophys. Res. Lett.* **45**, 4692–4700 (2018). doi: 10.1029/2018GL077925

33. G. Provan *et al.*, Planetary period oscillations in Saturn's magnetosphere: Cassini magnetic field observations over the northern summer solstice interval. *J. Geophys. Res. Space Phys.* **123**, 3859–3899 (2018). doi: 10.1029/2018JA025237

34. P. Galopeau, A. Lecacheux, Variations of Saturn's radio rotation period measured at kilometer wavelengths. *J. Geophys. Res.* **105** (A6), 13089–13101 (2000). doi: 10.1029/1999JA005089

35. D. A. Gurnett *et al.*, The reversal of the rotational modulation rates of the north and south components of Saturn kilometric radiation near equinox. *Geophys. Res. Lett.* **37**, L24101 (2010). doi: 10.1029/2010GL045796

36. G. Provan *et al.*, Planetary period oscillations in Saturn's magnetosphere: Coalescence and reversal of northern and southern periods in late northern spring. *J. Geophys. Res.* **121**, 9829–9862 (2016). doi: 10.1002/2016JA023056

37. X. Jia, M. G. Kivelson, T. I. Gombosi, Driving Saturn's magnetospheric periodicities from the upper atmosphere/ionosphere. *J. Geophys. Res.* **117** (A04215), 1460 (2012). doi: 10.1029/2011JA017367

38. X. Jia, M. G. Kivelson, Driving Saturn's magnetospheric periodicities from the upper atmosphere/ionosphere: Magnetotail response to dual sources. *J. Geophys. Res.* **117**, A11219 (2012). doi: 10.1029/2012JA018183

39. J. D. Anderson, G. Schubert, Saturn's gravitational field, internal rotation, and interior structure. *Science* **317**, 1384–1387 (2007). doi: 10.1126/science.1144835; pmid: 17823351

40. P. L. Read, T. E. Dowling, G. Schubert, Saturn's rotation period from its atmospheric planetary-wave configuration. *Nature* **460**, 608–610 (2009). doi: 10.1038/nature08194

41. R. Helled, E. Galanti, Y. Kaspi, Saturn's fast spin determined from its gravitational field and oblateness. *Nature* **520**, 202–204 (2015). doi: 10.1038/nature14278; pmid: 25807487

42. S. W. H. Cowley, E. J. Bunce, J. M. O'Rourke, A simple quantitative model of plasma flows and currents in Saturn's polar ionosphere. *J. Geophys. Res.* **109** (A5), A05212 (2004). doi: 10.1029/2003JA010375

43. G. J. Hunt *et al.*, Field-aligned currents in Saturn's southern nightside magnetosphere: Sub-corotation and planetary period oscillation components. *J. Geophys. Res.* **119**, 9847–9899 (2014). doi: 10.1002/2014JA020506

44. G. J. Hunt *et al.*, Field-aligned currents in Saturn's northern nightside magnetosphere: Evidence for inter-hemispheric current flow associated with planetary period oscillations. *J. Geophys. Res.* **120**, 7552–7584 (2015). doi: 10.1002/2015JA021454

45. B. A. Archinal *et al.*, Report of the IAU Working Group on Cartographic Coordinates and Rotational Elements: 2015. *Celestial Mech. Dyn. Astron.* **130**, 22 (2018). doi: 10.1007/s10569-017-9805-5

46. G. J. Hunt *et al.*, Field-aligned currents in Saturn's magnetosphere: Observations from the F-ring orbits. *J. Geophys. Res. Space Phys.* **123**, 3806–3821 (2018). doi: 10.1029/2017JA025067

47. K. K. Khurana *et al.*, Discovery of atmospheric-wind-driven electric currents in Saturn's magnetosphere in the gap between Saturn and its rings. *Geophys. Res. Lett.* (2018). doi: 10.1029/2018GL078256

48. M. E. Burton, M. K. Dougherty, C. T. Russell, Saturn's internal planetary magnetic field. *Geophys. Res. Lett.* **37**, L24105 (2010). doi: 10.1029/2010GL045148

49. D. J. Stevenson, Saturn's luminosity and magnetism. *Science* **208**, 746–748 (1980). doi: 10.1126/science.208.4445.746; pmid: 17771132

50. D. J. Stevenson, Reducing the non-axisymmetry of a planetary dynamo and an application to Saturn. *Geophys. Astrophys. Fluid Dyn.* **21**, 113–127 (1982). doi: 10.1080/03091928208209008

51. U. R. Christensen, J. Wicht, Models of magnetic field generation in partly stable planetary cores: Application to Mercury and Saturn. *Icarus* **196**, 16–34 (2008). doi: 10.1016/j.icarus.2008.02.013

52. S. Stanley, A dynamo model for axisymmetrizing Saturn's magnetic field. *Geophys. Res. Lett.* **37**, L05201 (2010). doi: 10.1029/2009GL041752

53. D. Gubbins, *Time Series Analysis and Inverse Theory for Geophysicists* (Cambridge Univ. Press, 2004).

54. R. Holme, J. Bloxham, The magnetic fields of Uranus and Neptune: Methods and models. *J. Geophys. Res.* **101**, 2177–2200 (1996). doi: 10.1029/95JE03437

55. P. Mauersberger, Das Mittel der energiedichte des geomagnetischen Hauptfeldes an der erdoberflasähe und seine säkulare Änderung. *Gerlands Beitr. Geophys.* **65**, 207–215 (1956).

56. F. J. Lowes, Mean-square values on sphere of spherical harmonic vector fields. *J. Geophys. Res.* **71**, 2179–2179 (1966). doi: 10.1029/JZ071i008p02179

57. F. J. Lowes, Spatial power spectrum of the main geomagnetic field, and extrapolation to the core. *Geophys. J. Int.* **36**, 717–730 (1974). doi: 10.1111/j.1365-246X.1974.tb00622.x

58. J. J. Liu, P. M. Goldreich, D. J. Stevenson, Constraints on deep-seated zonal winds inside Jupiter and Saturn. *Icarus* **196**, 653–664 (2008). doi: 10.1016/j.icarus.2007.11.036

59. The local magnetic Reynolds number is defined as $R_m = U_{conv}H_\sigma/\lambda$, where U_{conv} is the typical convective velocity; λ is the magnetic diffusivity, defined as the inverse of the product of electrical conductivity and magnetic permeability; and H_σ is the conductivity scale-height defined as $H_\sigma = \sigma/|d\sigma/dr|$.

60. U. R. Christensen, Dynamo scaling laws and application to the planets. *Space Sci. Rev.* **152**, 565–590 (2010). doi: 10.1007/s11214-009-9553-2

61. T. Gastine, J. Wicht, L. D. V. Duarte, M. Heimpel, A. Becker, Explaining Jupiter's magnetic field and equatorial jet dynamics. *Geophys. Res. Lett.* **41**, 5410–5419 (2014). doi: 10.1002/2014GL060814

62. G. A. Glatzmaier, Computer simulations of Jupiter's deep internal dynamics help interpret what Juno sees, *Proceedings of the National Academy of Sciences*, 201709125; DOI: 10.1073/pnas.1709125115 (2018).

63. Due the divergence free nature of magnetic field, a poloidal-toroidal decomposition of the magnetic field, $\mathbf{B} = \nabla \times (\nabla \times P\mathbf{e_r}) + \nabla \times T\mathbf{e_r}$, is commonly adopted in the dynamo region. The toroidal magnetic field, $\mathbf{B_T} = \nabla \times T\mathbf{e_r}$ has no radial component, and is associated with local electrical currents. The poloidal magnetic field $\mathbf{B_P} = \nabla \times (\nabla \times P\mathbf{e_r})$ has no azimuthal component if it is axisymmetric.

64. J. E. P. Connerney, M. H. Acuna, N. F. Ness, T. Satoh, New models of Jupiter's magnetic field constrained by the Io flux tube footprint. *J. Geophys. Res.* **103** (A6), 11929–11939 (1998). doi: 10.1029/97JA03726

65. Z. J. Yu, H. K. Leinweber, C. T. Russell, Galileo constraints on the secular variation of the Jovian magnetic field. *J. Geophys. Res.* **115** (E3), E03002 (2010). doi: 10.1029/2009JE003492

66. V. A. Ridley, R. Holme, Modeling the Jovian magnetic field and its secular variation using all available magnetic field observations. *J. Geophys. Res.* **121**, 309–337 (2016). doi: 10.1002/2015JE004951

67. K. M. Moore, J. Bloxham, J. E. P. Connerney, J. L. Jørgensen, J. M. G. Merayo, The analysis of initial Juno magnetometer data using a sparse magnetic field representation. *Geophys. Res. Lett.* **44**, 4687–4693 (2017). doi: 10.1002/2017GL073133

68. Assuming 1 cm s^{-1} flow and an electrical conductivity of 2×10^5 S/m, R_m associated with a layer with a thickness of 0.2 R_S would be ~30,000.

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIALS

RESEARCH ARTICLE SUMMARY

GAS GIANT PLANETS

A radiation belt of energetic protons located between Saturn and its rings

E. Roussos*†, P. Kollmann†, N. Krupp, A. Kotova, L. Regoli, C. Paranicas, D. G. Mitchell, S. M. Krimigis, D. Hamilton, P. Brandt, J. Carbary, S. Christon, K. Dialynas, I. Dandouras, M. E. Hill, W. H. Ip, G. H. Jones, S. Livi, B. H. Mauk, B. Palmaerts, E. C. Roelof, A. Rymer, N. Sergis, H. T. Smith

INTRODUCTION: Most magnetized planets are known to possess radiation belts, where high-energy charged particles are trapped in large numbers. The possibility that a radiation belt could exist also in the confined region between Saturn and its main rings has been proposed on the basis of remote sensing observations and simulations. It was not until the final 5 months of the Cassini mission that in situ measurements were obtained from this region with the Magnetosphere Imaging Instrument (MIMI). This paper provides an overview of these measurements and their interpretation.

RATIONALE: Saturn's main rings prevent the inward transport of trapped charged particles in the magnetosphere. Material from the outer radiation belts cannot directly access the low-altitude region within the rings. The isolation of this region allows the study of energetic particle source and loss processes because it is only indirectly coupled to the dynamics of the rest of the magnetosphere. Potential sources include cosmic ray albedo neutron decay (CRAND) and multiple-charge exchange, whereas losses are likely dominated by energy deposition and scattering of trapped particles by dust and atmospheric neutrals. All of these mechanisms involve charged particle interactions with materials in space, meaning that MIMI measurements can provide information to probe the material itself—particularly the tenuous D-ring, the innermost component of Saturn's main rings, which is difficult to constrain by remote sensing observations.

RESULTS: We observed an inner radiation belt extending between 1.03 and 1.22 Saturn radii (1 R_S = 60,268 km) at the equatorial plane, dominated by protons with energies from 25 MeV up to the giga–electron volt range. This belt is limited by the atmosphere at its inner edge and by the D73 ringlet (at 1.22 R_S), a component of the D-ring, at its outer boundary. Another ringlet (D68 at 1.12 R_S) splits the trapped particle population in two. The outer sector overlaps with the extended D-ring, and its intensity is reduced compared with that of the inner sector, owing to proton losses on ring dust. The proton angular distributions are highly anisotropic with fluxes that are orders of magnitude higher near the magnetic equator compared with fluxes of particles that can reach high latitudes. No time variability could be discerned in the >25-MeV proton population over the 5-month period of the observations. Trapping of lower-energy (tens of kilo–electron volt) protons was clearly observed in at least one case by imaging the emission of energetic neutral atoms (ENAs) coming from below ~1.06 R_S (altitude < 3800 km). Energetic electrons (18 keV to several mega–electron volts) and heavy ions (27 keV per nucleon to hundreds of mega–electron volts per nucleon), if present, have fluxes close to or lower than the detection limit of the MIMI sensors.

CONCLUSION: The radial profile, the stability of the >25-MeV proton fluxes, and the lack of heavy ions are features consistent with a radiation belt originating from CRAND. The strong anisotropy of the proton distributions is primarily the result of proton losses in collisions with atmospheric neutrals, though an anisotropy in the production of CRAND protons from Saturn's rings may also contribute. The low-altitude, kilo–electron volt proton population is transient and derives from charge stripping of planetward ENAs, which are generated at the variable magnetospheric ring current. ∎

Saturn's proton radiation belts. Saturn's permanent proton radiation belt extends outward to the orbit of the moon Tethys but is segmented because of proton absorption by moons and rings. The innermost radiation belt (inset) threads through Saturn's D-ring and contains protons with energies up to several giga–electron volts, much higher than observed outside the main rings. These protons are among the β-decay products of neutrons, which are released through galactic cosmic ray collisions with Saturn's rings (CRAND process).
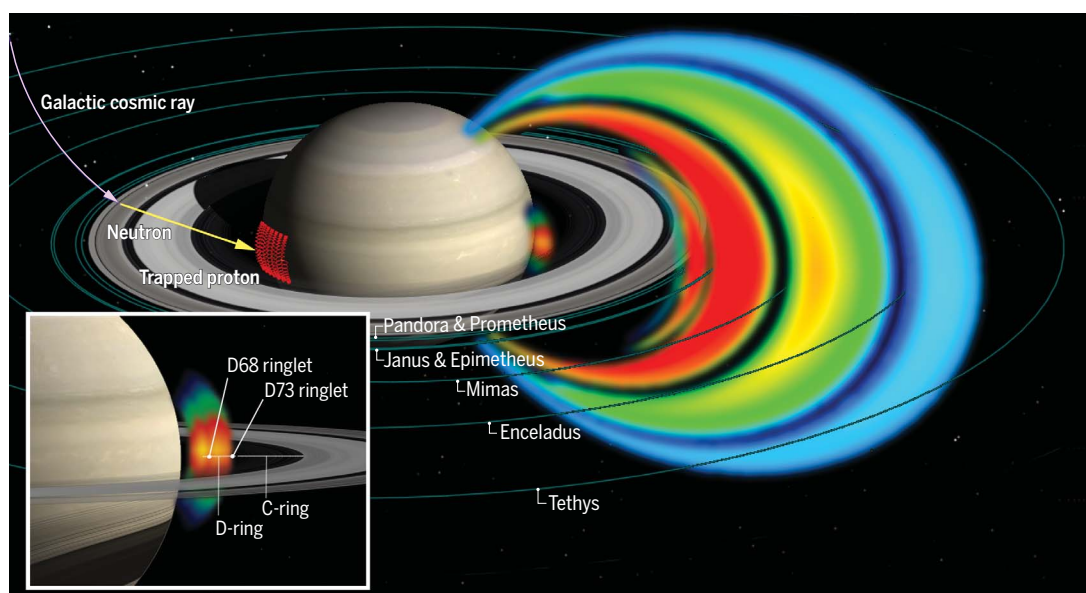
RESEARCH ARTICLE

GAS GIANT PLANETS

A radiation belt of energetic protons located between Saturn and its rings

E. Roussos[1]*†, P. Kollmann[2]†, N. Krupp[1], A. Kotova[3], L. Regoli[4], C. Paranicas[2],
D. G. Mitchell[2], S. M. Krimigis[2,5], D. Hamilton[6], P. Brandt[2], J. Carbary[2], S. Christon[7],
K. Dialynas[5], I. Dandouras[3], M. E. Hill[2], W. H. Ip[8], G. H. Jones[9,10], S. Livi[11],
B. H. Mauk[2], B. Palmaerts[12], E. C. Roelof[2], A. Rymer[2], N. Sergis[5,13], H. T. Smith[2]

Saturn has a sufficiently strong dipole magnetic field to trap high-energy charged particles and form radiation belts, which have been observed outside its rings. Whether stable radiation belts exist near the planet and inward of the rings was previously unknown. The Cassini spacecraft's Magnetosphere Imaging Instrument obtained measurements of a radiation belt that lies just above Saturn's dense atmosphere and is decoupled from the rest of the magnetosphere by the planet's A- to C-rings. The belt extends across the D-ring and comprises protons produced through cosmic ray albedo neutron decay and multiple charge-exchange reactions. These protons are lost to atmospheric neutrals and D-ring dust. Strong proton depletions that map onto features on the D-ring indicate a highly structured and diverse dust environment near Saturn.

During the proximal orbit phase of the Cassini mission (23 April to 15 September 2017), the spacecraft completed 22 crossings through the narrow gap between Saturn's upper atmosphere and its rings (Fig. 1). These orbits provided the opportunity to conduct in situ measurements of the local energetic charged particle environment with the Magnetosphere Imaging Instrument (MIMI) (1).

Observations of trapped particle radiation confined inward of Saturn's rings (inner trapping region) were obtained during Cassini's Saturn Orbit Insertion (SOI; 1 July 2004) through energetic neutral atom (ENA) imaging of this region using MIMI (2). Those observations revealed the emission of ENAs with energy of 20 to 50 keV per nucleon from a low-altitude, trapped ion population of the same energy, which was subject to

charge-exchange with neutral atoms of Saturn's upper atmosphere. The ions producing the low-altitude ENA emission are thought to derive from ENAs generated in the planet's middle magnetosphere that propagate toward Saturn and impact its atmosphere (3, 4). Following their reionization through charge-stripping reactions in the planet's upper atmosphere, newly converted ions are temporarily trapped by the planet's magnetic field before charge exchange converts them back into the ENAs detected by MIMI (2). The altitude of the ion population driving this ENA emission is unknown but could extend to altitudes above Saturn's dense atmosphere (>1300 km), similar to the 0.06- to 1-MeV ions that were recently detected between 4300 and 18,000 km above Jupiter's 1-bar atmospheric level and were possibly generated by the same mechanism (multiple charge-exchange reactions) (5).

Higher-energy protons, at mega–electron volt energies and above, could be supplied to the inner trapping region through the cosmic ray albedo neutron decay (CRAND) process (6, 7). CRAND protons are among the β-decay products of secondary (albedo) neutrons formed in galactic cosmic ray (GCR) impacts on Saturn's atmosphere and/or its dense rings. Because neutrons are not bound by the planetary magnetic field, they can propagate from their generation site and through the trapping region, within which they may release their β-decay protons. The process is known to sustain the proton radiation belts of Earth and those of Saturn outside of its rings (main radiation belts) (8–12).

Although CRAND generates energetic protons near the planet, it has not been clear whether these particles can accumulate in large numbers and form a localized radiation belt. Quantitative models used to predict upper limits on the

proton fluxes in this region (13) relied on a series of input parameter extrapolations and simplifying assumptions to determine the CRAND source rate and the loss rates of protons to atmospheric neutrals and to ring dust. It was not known if there would be a detectable signal from energetic protons across the D-ring (~1.11 to 1.24 R_S, where 1 R_S = 60,268 km is Saturn's radius) (14) and its three ringlets—named D68, D72, and D73 and centered at 1.12, 1.19, and 1.22 R_S, respectively (15, 16)—all of which are contained within the inner trapping region. The D-ring properties were so poorly constrained that predictions for the 10- to 60-MeV proton fluxes spanned more than two orders of magnitude (13), with the lowest values near MIMI's expected detection limits.

The simulations indicated that the dynamics of mega–electron volt protons in this inner radiation belt would be determined by different physical processes than those that affect the main radiation belts. Proton fluxes in Saturn's main radiation belts are limited by radial diffusion, which controls how fast these particles are distributed within the orbits of the planet's large icy moons, where they are subsequently absorbed (12, 17). Similarly, radial diffusion near the planet could act as a proton sink by gradually moving protons to the massive C-ring (1.24 to 1.53 R_S) and the dense atmosphere ($\leq$ 1.02 R_S). However, radial diffusion rates near the planet were projected to be extremely low (17, 18), so it was proposed that the intensity of the proton fluxes would be primarily determined by a balance of the CRAND source rate against losses of protons to dust and neutral gas. Losses to the equatorially confined dust can be transmitted along the magnetic field lines sampled by Cassini's high-inclination proximal orbit trajectory, allowing MIMI to obtain radial dust density scans of the D-ring system from a large distance (Fig. 1).

Understanding how the local dust and gas environment interacts with CRAND protons is simplified because proton populations in the inner trapping region are permanently isolated from the rest of the magnetosphere. Saturn's dense A- to C-rings form a ≈62,000-km-wide particle-absorbing corridor (1.24 to 2.27 R_S) that is impermeable to any magnetospheric particle transported inward of 2.27 R_S. Mixing of energetic particle populations near the planet from different source locations (10), which may occur at Earth and Jupiter and may complicate the interpretation of relevant measurements, is inhibited at Saturn. This filtering also applies to CRAND electrons, the other β-decay product of albedo neutrons, which were only recently detected around Earth (19). Furthermore, the strong magnetic field near Saturn forms a stable energetic charged particle trapping environment that is not perturbed by solar wind or magnetospheric transients that affect the magnetosphere at larger radii. The location of the inner trapping region establishes a constant source rate for CRAND protons and electrons because the primary GCR energies that drive CRAND and define its source strength exceed 20 GeV (20). GCRs

[1]Max Planck Institute for Solar System Research, 37077 Goettingen, Germany. [2]Johns Hopkins University Applied Physics Laboratory (JHUAPL), Laurel, MD 20723, USA. [3]Institut de Recherche en Astrophysique et Planétologie, Université de Toulouse, Centre National de la Recherche Scientifique, Université Paul Sabatier, Centre National d'Etudes Spatiales, Toulouse, France. [4]Department of Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor, MI 48109, USA. [5]Office of Space Research and Technology, Academy of Athens, Athens 11527, Greece. [6]Department of Physics, University of Maryland, College Park, MD 20742, USA. [7]Focused Analysis and Research, Columbia, MD 11043, USA. [8]Institute of Astronomy, National Central University, 32001 Jhongli, Taiwan. [9]Mullard Space Science Laboratory, University College London, Dorking, Surrey RH5 6NT, UK. [10]The Centre for Planetary Sciences at University College London/ Birkbeck, London WC1E 6BT, UK. [11]Southwest Research Institute, San Antonio, TX 78238, USA. [12]Laboratory for Planetary and Atmospheric Physics, Space Sciences, Technologies and Astrophysics Institute, University of Liege, Liege, Belgium. [13]National Observatory of Athens, 15236 Penteli, Greece.
*Corresponding author. Email: roussos@mps.mpg.de
†These authors contributed equally to this work.

above this energy are not affected by the solar cycle ([21]). The stability of the CRAND source rate has previously proven useful in the understanding of the magnetospheric processes that control the evolution of Saturn's main radiation belts ([12]).

Instrumentation and methodology

The MIMI instrument ([1]) comprises three different sensors: the Low Energy Magnetospheric Measurement System (LEMMS), the Charge-Energy-Mass Spectrometer (CHEMS), and the Ion Neutral Camera (INCA).

LEMMS is a double-ended, charged particle telescope that can measure the energy and angular distribution of 27-keV to >300-MeV protons and of 18-keV to ~10-MeV electrons ([22]). Some LEMMS channels can also distinguish heavier mega–electron volt ions from protons but lack mass resolution ([23]). CHEMS can measure the energy, mass, and charge state of energetic ions between 3 and 220 keV per electron (e^-). INCA obtains ENA images of oxygen and hydrogen and high-sensitivity ion spectra in the energy range from 7 keV per nucleon to 8 MeV per nucleon.

The majority of our results are based on observations by LEMMS channels P8 (>25 MeV H^+) and E7 (>300 MeV H^+ and >7 MeV e^-). These two channels achieve the most-efficient rejection of instrument-penetrating protons that can contaminate the measurements. We occasionally use channel P9 (>60 MeV H^+ and >1 MeV e^-) because of its high-sensitivity, omnidirectional proton response. Unless otherwise stated, we quote channels P8, E7, and P9 by their proton energy response, because the respective measurements are dominated by protons.

Differential proton flux spectra were obtained using a forward model that reconstructs the >25- and >300-MeV count rates by convolving those channels' angular and energy response functions with predefined proton energy spectra and angular distributions ([22]) (figs. S2 and S6). We also detected upper limits for lower proton energies or nonproton species when no signal was detected in LEMMS, CHEMS, and INCA ([22]) (fig. S7).

Magnetic mapping of the in situ measurements was performed by calculating the L-shell, equatorial pitch angle, and the loss cone through an empirical third-order magnetic field model ([24]). We define the L-shell value (L) as the distance from the planet (in Saturn radii) that a field line crosses the magnetic equator. The equatorial pitch angle (α_{eq}) is the angle between the proton velocity and the magnetic field at the magnetic equator, whereas the loss cone corresponds to the pitch angle below which the trapped particles reach under an altitude of 1000 km and are absorbed in the dense atmosphere.

Inner trapping region observations
Raw proton data

The raw count rates of >25-, >300-, and >60-MeV protons are plotted in Fig. 2 as a function of L, from all instances when Cassini was magnet-

ically connected to regions inward of Saturn's C-ring (table S1) ([22]). We can identify several features in this L-shell profile even before converting raw count rates to physical units.

We find a strong signal on magnetic field lines that map to the D-ring, even though trapped energetic protons at those L-shells cross through that ring every few seconds during their oscillation (bounce motion) along the magnetic field lines and thus are expected to experience heavy losses. The effects of proton absorption from two of the D-ring's ringlets are more severe. The count rate dropout that develops at the outer boundary of the trapping region maps to the location of the D73 ringlet (which has a width of 0.02 R_S), not the C-ring. This makes the radiation belt slightly narrower than predicted, because the presence of the ringlets was not taken into account in past simulations ([13]). A second drop in count rate, which splits the inner radiation belt into two main segments, is seen at the L-shell of the D68 ringlet. No obvious absorption signature is observed in association with the D72 ringlet.

In agreement with model predictions ([13]), the dense upper atmosphere limits the fluxes of trapped radiation toward the lowest L-shells. The omnidirectional >60-MeV proton measurements (Fig. 2C) registered counts at levels above the instrumental background down to $L \sim 1.03$, indicating that some minimal flux of mega–electron volt protons survives to at least 1800 km above the 1-bar atmospheric pressure level.

Proton pitch angle distributions

The large scatter of the mega–electron volt proton rates observed at any given L-shell (Fig. 2) can be attributed to changes in both the spacecraft's latitude and LEMMS's equatorial pitch angle pointing, α_{eq} (fig. S1) ([22]). Reconstructing these two dominant dependencies (fig. S6) ([22]) requires the proton pitch angle distribution (PAD) to be much steeper than the one LEMMS resolves with its coarse angular resolution. If we describe the PAD as $\propto \sin^N \alpha_{eq}$ outside of the loss cone, we obtain that the power, N, ranges from 10 (at the D-ring) to 100 (near the atmosphere).
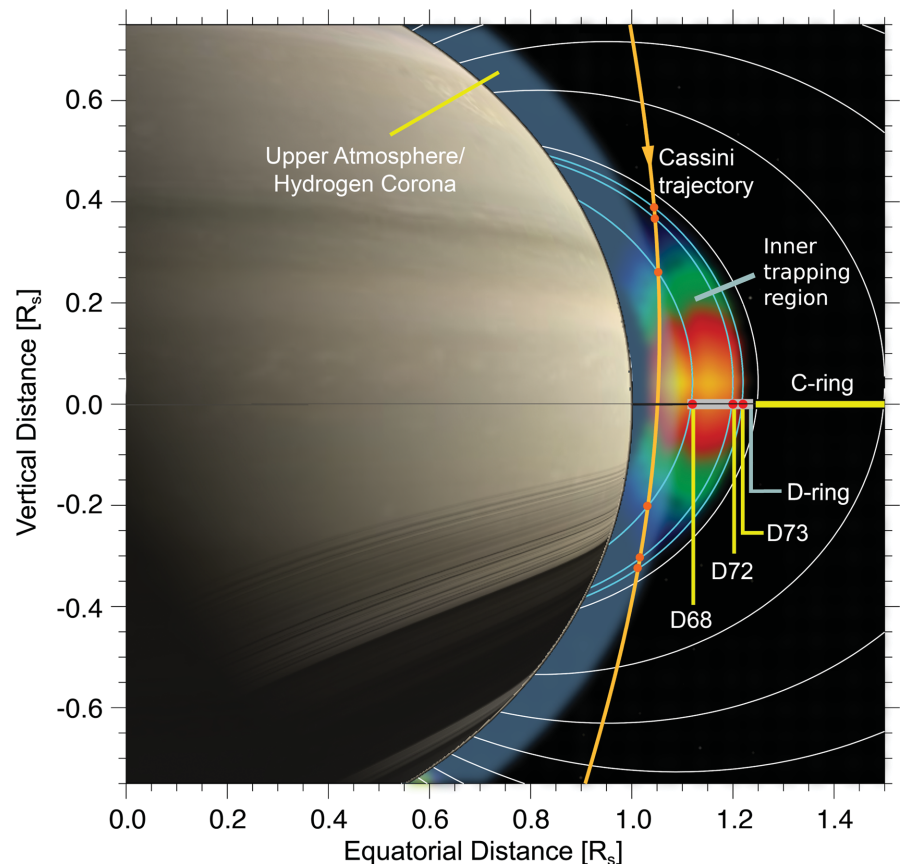

Fig. 1. Geometry of the inner trapping region between Saturn's rings and atmosphere. A schematic illustration of the inner trapping region is indicated by the color map. A typical Cassini proximal orbit trajectory is shown in orange. Cassini first intersects this region at ~20° north latitude and exits at a similar latitude to the south. Because trapped particles oscillate (bounce) along the magnetic field and through the equator every 2 to 3 s, Cassini can probe the effects of the D-ring's dust environment on these particles from high latitudes without directly crossing through it. The orange circles mark the locations along Cassini's trajectory where the effects of the D-ring's ringlets (D68, D72, and D73) on the trapped particle population can be can be transmitted along the magnetic field lines drawn in light blue. Other field lines (in white) are shown every 0.25 R_S at the equatorial plane.

For comparison, in Saturn's main proton radiation belts, this exponent is always less than ~6 (25–27). The small residuals between the observations and the reconstruction indicate that temporal variations, if present, must be less than the 1σ statistical uncertainty in the LEMMS signal.

Proton energy spectra (≥25 MeV)

Differential proton fluxes were evaluated, assuming a spectral form that is a simple power law in energy and has a cutoff at 20 GeV. Even if Saturn's magnetic field can stably trap even higher energies (28), those protons' circular motion around the magnetic field lines has a radius (gyroradius) that is comparable to the width of the trapping region (29). The resulting spectrum at L = 1.1, where the proton fluxes peak, is plotted in Fig. 3A.

We find that the proton spectrum is hard (i.e., flux decreases slowly with increasing energy), with a spectral index of ~–1. For such a spectrum, relativistic protons ($\gtrsim$ 0.9 GeV) could make a considerable contribution to the count rate of the >300-MeV channel. The fact that the >300-MeV proton channel recorded counts even when LEMMS was pointing well into the atmospheric loss cone (Fig. 2, A and B, and fig. S1A) is

an indirect but independent verification that relativistic protons contribute to the measurements: Such counts may only come from >0.9-GeV instrument-penetrating protons (fig. S3) (22).

The shape of the proton spectrum may be more complex than the power law we assumed. For example, an additional spectral break might be expected at ~100 MeV, where the efficiency of CRAND neutron production changes (9, 30). Such a spectral shape cannot be unambiguously constrained with just the two proton channels used in this study. Still, even if we predefine its shape, the spectrum at >100 MeV remains hard, meaning that it is possible to constrain the proton fluxes at the giga–electron volt range. Our estimation of fluxes at 1 GeV is shown in the proton spectrum of Fig. 3A.

L-shell distribution of protons

The L-shell dependence of the differential proton fluxes at 300 MeV, deconvolved from their latitudinal and pitch angle dependencies (Fig. 4), shows that the innermost proton belt peaks around L = 1.1, just inward of the inner D-ring edge and in agreement with model predictions (13). That is a region where the combined material density from the D-ring dust and the atmo-

spheric neutrals reaches a minimum, raising the proton fluxes.

The depth of the dropout attributed to the D68 ringlet has a pitch angle dependence (Fig. 4). With decreasing α_{eq}, the depth of the absorption becomes less pronounced, as expected for charged particles reaching high latitudes as they bounce along the magnetic field lines, well away from where proton losses to dust occur. No absorption signature is resolved at the location of the D72 ringlet, even after the raw counts are processed and deconvolved from their latitudinal and pitch angle dependencies.

Low-altitude ENA emissions and upper limits of <25-MeV proton, ion, and electron fluxes

Upper limits for <25-MeV protons, 18- to 832-keV electrons, and >5-MeV-per-nucleon helium and oxygen are also shown in Fig. 3, A and B. The results apply to L = 1.1 but are similar (within a factor of 3) to the entire L-shell range of the inner trapping region. The upper limits indicate a drop of the proton fluxes below ~25 MeV and negligible fluxes for light or heavy ions. Evidence for the absence of detectable electron fluxes at any energy to which LEMMS responds (18 keV to ~10 MeV) is shown in figs. S4 and S5 (22). The only resolvable trapped particle population within situ data comes from >25-MeV protons.

Even though MIMI did not observe kilo–electron volt protons in situ, it detected remotely a 24- to 55-keV proton ENA emission from Saturn during its closest approach to the planet on 28 May 2017 (Fig. 5). The proton population responsible for this emission must have resided below the lowest altitude of ~3800 km sampled in situ by MIMI on that day. The non-detection of ENAs in several other cases that INCA had the correct pointing to observe (e.g., on 2 September 2017) suggests that the protons creating the ENA emission are transient.

Interpretation of observations by MIMI
Origin of mega–electron volt protons from ring CRAND

Several lines of evidence demonstrate that CRAND is the primary source process of the >25-MeV proton belt: the presence of protons, the extension of the spectrum well above 300 MeV, the lack of any resolvable signal from heavy ions, the temporal stability, and the L-shell profile of the proton fluxes. The L-shell profile (Fig. 4) agrees qualitatively with that derived from simulations (13) for which a CRAND source was used as an input. Because this inner radiation belt is permanently decoupled from the rest of the magnetosphere, its detection alone constitutes direct evidence of the CRAND process.

At Saturn, CRAND may be catalyzed not only through the planet's atmosphere but also through the rings, from which we argue that the majority of CRAND protons originate. One reason is that the rings have a higher neutron yield than the planet's atmosphere (30). In addition, atmospheric neutrons that can reach the inner trapping region may only originate from a latitude below 36°, the region accessible to >40-GeV GCRs

Fig. 2. MIMI/LEMMS count rates as a function of magnetic L-shell. Data are obtained from all proximal orbits (23 April to 15 September 2017) for regions that magnetically map inward of Saturn's C-ring (table S1) (22). All channels shown are dominated by protons and cover the energy range above 25 MeV. Count rates are averages of three consecutive samples (one data point per ~16 s). The points are color coded according to the percentage of LEMMS's aperture within the planetary loss cone (see legend). The channels in (**A**) and (**B**) roughly organize with pitch angle, whereas the channel in (**C**) does not, owing to its high sensitivity to sideways, instrument-penetrating protons.
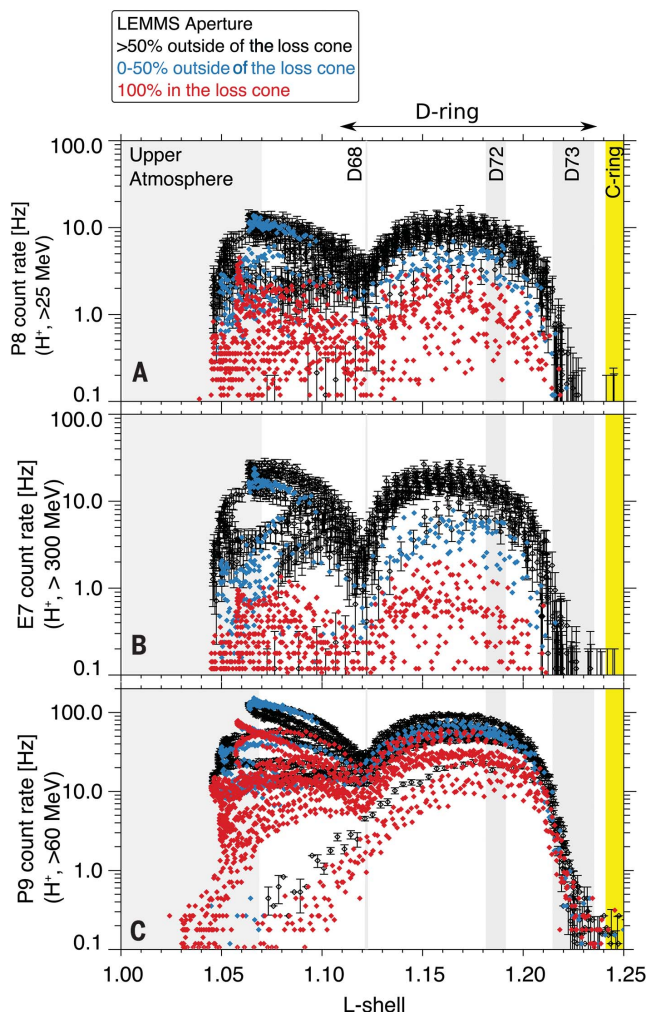
(20). Such GCRs have about one-fourth as much integral flux as the >20-GeV GCRs that reach the main rings, which also offer a 50% larger neutron production area than the limited atmospheric zone. The attribution of the inner radiation belt source to ring CRAND allows the separation of atmospheric and ring CRAND in the proton spectra of Saturn's main radiation belts (12, 17).

Radiation belt and atmosphere coupling

The steep PAD inferred for both >25- and >300-MeV protons inward of the D-ring ($L \lesssim 1.1$) can be attributed to energy losses of these protons to Saturn's extended atmosphere. We demonstrate this in Fig. 4 by plotting the inverse value of the average atmospheric density encountered by protons of different pitch angles as they move along the magnetic field (bounce-averaged atmospheric density) against the fluxes of 300-MeV protons. The inverse density profiles track the drop of the proton fluxes toward the planet, mostly for $L \lesssim 1.1$. In a similar way, the terrestrial atmosphere is responsible for very steep proton PADs in Earth's radiation belts (31, 32).

Radiation belt and D-ring coupling

For $L \gtrsim 1.1$, the inverse atmospheric density curves deviate from the deduced 300-MeV proton flux profiles, suggesting that the losses to the D-ring develop faster there. Using the previously derived scaling between D-ring density to proton flux (13), we estimate that the D-ring column density must be below 10^{-8} g/cm² (13), which is orders of magnitude lower than the corresponding A- and B-ring values [10 to 500 g/cm² (33–36)]. The partial, dust-driven depletion of the proton fluxes that we observe resembles the mega–electron volt proton interaction seen at Saturn's G-ring ($L = 2.71$) (17, 37, 38). Because losses to D-ring dust are stronger for equatorially mirroring particles ($\alpha_{eq} \sim 90°$), they are likely responsible for the reduced pitch angle anisotropy of protons estimated for $L \gtrsim 1.1$ ($N \sim 10$) compared with the anisotropy for $L \lesssim 1.1$ ($N \sim 100$).

Even though the anisotropy for $L \gtrsim 1.1$ is relatively small compared with that for $L < 1.1$, it is still large compared with the anisotropy of protons in Saturn's magnetosphere ($N \lesssim 6$). An additional source of proton anisotropy that could maintain such large N values may come from the CRAND process: Because CRAND protons are injected along the direction of their parent, β-decay neutrons, their PAD may retain information about a preferential emission direction of neutrons from the rings. Simulations show that an isotropic, ring neutron emission would lead to similarly isotropic proton PADs near the planet (18). The proton PAD at $L \gtrsim 1.1$ may then hold evidence that the neutron injection from the rings is highly anisotropic. Such an anisotropy may constrain the dust size distribution in Saturn's A- to C-rings, as has been previously suggested (9, 30).

Diversity of the D-ring ringlets

Only two (D68 and D73) of the three ringlets of the D-ring are found to influence the L-shell

profile of the proton fluxes. D73 has a normal optical depth of ~10⁻³ (14), sufficient to deplete the energetic protons that LEMMS detects (39). D68 causes a strong reduction of proton fluxes (Figs. 2 and 4), whereas D72, which is as optically bright as D68, appears to have no impact on the trapped protons. The different influences of D68 and D72 on protons suggest that the former ringlet concentrates more mass in large grains and/or in its longitudinally confined arc, which has been observed remotely (14).

Comparison with Earth's radiation belts

Figure 3A shows the >25-MeV proton spectra from Saturn's inner trapping zone compared with those from $L = 1.4$ in Earth's magnetosphere, where the proton belt fluxes above 100 MeV peak (40). The comparison indicates that Earth fluxes are about an order of magnitude higher for energies below 400 MeV. This difference is mostly due to solar protons, which can reach low L-shells at Earth through radial transport, a source not available at Saturn due to the A- to C-rings. A turnover occurs beyond ~400 MeV

and into the giga–electron volt range, where the projected proton fluxes at Saturn become stronger. Even though both Earth and Saturn produce CRAND protons in the giga–electron volt range, the 600-fold stronger magnetic moment at Saturn allows a much more stable trapping of protons at relativistic energies (28). At Earth, trapped protons have been observed up to 2.2 GeV (32), whereas stable proton trapping is estimated to extend up to ~5 GeV (41).

Comparison with Saturn's main radiation belts

Proton fluxes in Saturn's main radiation belts drop more steeply with increasing energy (11, 23, 42) compared with our results for the innermost belt. Several explanations could account for this difference. Atmospheric CRAND from high latitudes, which can be generated by high-flux >0.5-GeV primary GCRs (20), has a stronger contribution to the main radiation belts than the contribution to the radiation belt near the planet. Magnetospheric radial diffusion becomes increasingly important at larger L-shells (12),
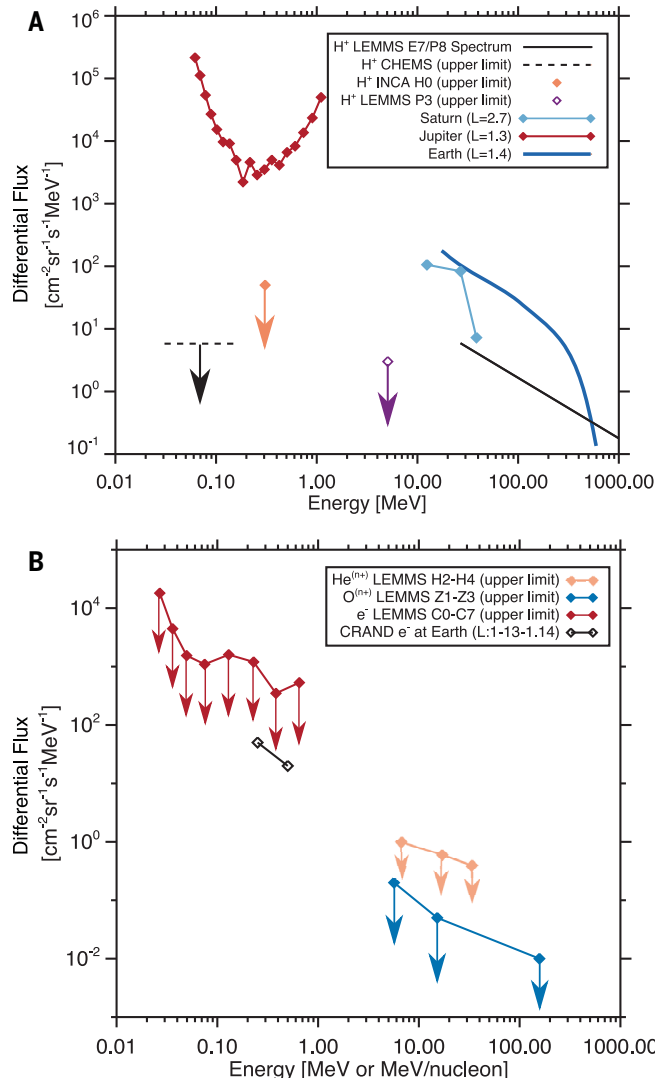

Fig. 3. Energetic particle spectra and upper limits. (**A**) Proton differential flux spectrum of 25-MeV to >1-GeV protons (black line). As the PAD of the 25-MeV to ~1-GeV protons is highly anisotropic, a pitch-angle–averaged spectrum is plotted. All other symbols denote upper flux limits for lower-energy protons. For comparison, we also show a proton spectrum from $L = 1.3$ at Jupiter (5), from $L = 2.7$ at Saturn (11), and from $L = 1.4$ ($\alpha_{eq} \sim 90°$) at Earth (40). (**B**) Flux upper limits for ions and electrons. The H2-H4 and Z1-Z3 channels constrain the minimum atomic number (Z) of the ions detected ($Z \geq 2$ and $Z \geq 8$, respectively), but, to define the channels' energies, we assume that these ions are helium and oxygen. On the horizontal axes, units are in mega–electron volts for protons and electrons and in mega–electron volts per nucleon for ions. The upper limits of electron fluxes can be compared with the maximum electron CRAND fluxes measured at Earth (19). All Saturn innermost belt measurements are from $L = 1.1$.

whereas the proton trapping limit also drops to the range of ~1 GeV. We cannot exclude the possibility that part of the difference seen in the spectra is due to differing LEMMS calibration between the studies.

Electron CRAND and additional energetic particle sources

We did not detect any signature of electrons (presumably from CRAND) below 837 keV. This suggests that any electrons produced by CRAND are lost more efficiently than protons and cannot build up detectable fluxes. The CRAND source rates at Earth and Saturn are comparable (6), so we can use this nondetection and make a rough order-of-magnitude estimation for the time scales of CRAND electron losses in the inner belt.

If we assume that the input rate to 500-keV CRAND electron fluxes is 2.5×10^{-2}/(keV cm^2 sr s) in 1.5 hours, as measured at Earth (19), we find that CRAND electrons would exceed the upper detection limits shown in Fig. 3B within just 1 to 2 days. That is much shorter than the expected, year-long trapping time scales in the strong, axisymmetric magnetic field near Saturn. The nondetection indicates that 18- to 837-keV electrons are subject to losses, which act faster than a few days and could develop from electron scattering due to dust and neutrals, from wave-particle interactions, or even from weak radial flows, to which electrons are much more sensitive than the protons (43). Such flows may drive CRAND electrons onto the C-ring or the atmosphere within a few hours or days after their injection into the trapping region.

The upper limits for heavy mega–electron volt ions set tight constraints on the intensity of other, non-CRAND related source process that may operate in this region, such as the local production of energetic light ions through elastic collisions of CRAND protons with atmospheric neutrals (44).

Low-altitude, kilo–electron volt proton radiation belt

The transient character of the low-altitude ENA emission confirms that its origin is in the variable ring current (2, 45). The 3800-km altitude limit inferred based on the lack of in situ kilo–electron volt proton detection when the ENA emission image was obtained (Fig. 5) can be further reduced to 2700 km if we rely on the value of the lowest L-shell for which LEMMS had the appropriate pointing to observe these protons in situ.

We find numerous differences when we compare our findings to similar observations at Jupiter. There, kilo–electron volt protons, which may have the same origin (charge-stripped ENAs), have fluxes that are three to five orders of magnitude above the upper limits estimated for Saturn (5) (Fig. 3A). It is possible that the highly structured internal magnetic field of Jupiter (46) enables ions produced through charge stripping in the denser layers of its upper atmosphere to drift around the planet along paths that extend to higher altitudes, where the bounce-averaged density of neutrals to which they are exposed becomes negligible. That could allow ion fluxes to accumulate and become detectable before the ions complete a full orbit around the planet. The nonaxisymmetric terrestrial magnetic field has a similar effect on CRAND electrons, as it allows them to accumulate only at a restricted longitude range where their orbit does not intersect the dense layers of our planet's atmosphere (19). The axisymmetric magnetic field of Saturn, on the other hand, restricts ions produced from charge stripping of ENAs to their high–atmospheric density generation altitude, severely limiting their lifetime and the extension of their population to the L-shells where MIMI made its in situ observations.

Summary and conclusions

The proximal orbits of Cassini allowed the MIMI instrument to sample unexplored regions of Saturn's magnetosphere. Our measurements demonstrate that a radiation belt sector is present inward of Saturn's dense rings, despite the isolation of this region from the rest of the magnetosphere and its collocation with dust and atmospheric neutrals.

This radiation belt has two components. The primary one, bound by the planet's atmosphere and the D73 ringlet, originates from ring CRAND, comprises protons with energies extending from
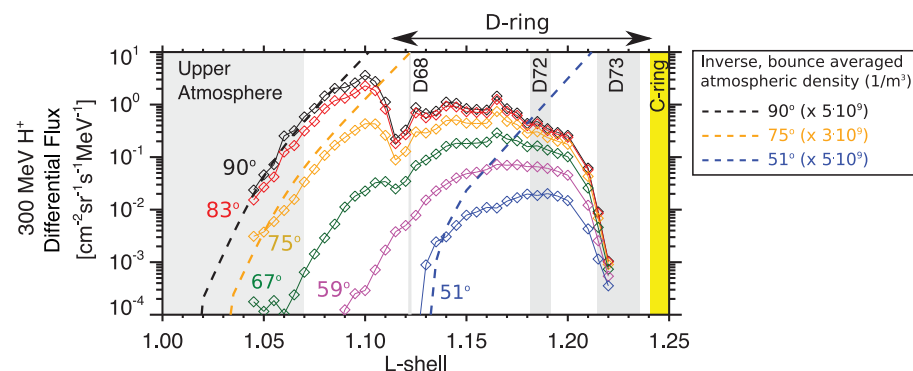

Fig. 4. L-shell dependence of 300-MeV proton differential fluxes. Different lines and colors correspond to different α_{eq}, as indicated in the figure. Fluxes are shown only where the inversion was successful. The progressively smaller extension of the radiation belts for decreasing pitch angles is due to the increasing size of the loss cone toward lower L-shells. The small offset of the central dropout from the L-shell of D68 is probably due to systematic errors associated with the magnetic field model. Overplotted are three curves showing the inverse of the bounce-averaged atmospheric density (13), scaled by appropriate factors (see legend) for a better comparison with the proton fluxes. The small difference in the adopted scaling factors accounts for the inconsistencies between the proton flux profiles and the atmospheric model used.

Fig. 5. Imaging of the low-altitude ENA emission. An image of Saturn in 24- to 55-keV proton ENAs obtained with the MIMI/INCA camera on 28 May 2017 is shown. Cassini was 1.13 R_S away from Saturn's center and at a north latitude of 20.45°. The grid of Saturn is overlaid for reference, with the red grid line marking the equator. Latitude and longitude grid lines are shown every 10°. The emission from the low-altitude kilo–electron volt ion radiation belt, observed also at SOI in 2004 (2), is the one from lower latitudes. The emission that appears to come from higher latitudes is due to contamination of the INCA measurements by the >25-MeV protons observed by LEMMS. A smoothing has been applied to the image to enhance visibility of the low-altitude emission and to reduce signal fluctuations due to low counting statistics. The boundaries of the emission are due to INCA's field of view. The coordinate system displayed has its origin at the center of Saturn, with the z axis pointing north. The other two axes are along Saturn's equatorial plane, with the y axis pointing toward dusk and the x axis toward local noon.
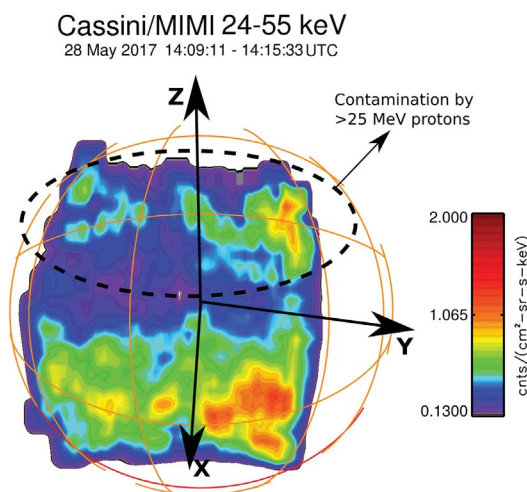
~25 MeV into the giga–electron volt range, and appears to be stable throughout the 5-month observation period. Its structure is modified by the interaction of the CRAND protons with Saturn's atmosphere, the D-ring and its three ringlets. The secondary, low-altitude radiation belt component was observed only remotely through ENA imaging, is transient, and contains protons of kilo–electron volt energies. The nondetection of the kilo–electron volt proton population responsible for the ENA emission of the secondary belt, in the context of equivalent measurements from Jupiter and Earth, shows that Saturn's axisymmetric magnetic field suppresses the formation of stable, low-altitude radiation belts through multiple-charge exchange.

REFERENCES AND NOTES

1. S. M. Krimigis et al., Magnetosphere Imaging Instrument (MIMI) on the Cassini Mission to Saturn/Titan. *Space Sci. Rev.* **114**, 233–329 (2004). doi: 10.1007/s11214-004-1410-8
2. S. M. Krimigis et al., Dynamics of Saturn's magnetosphere from MIMI during Cassini's orbital insertion. *Science* **307**, 1270–1273 (2005). doi: 10.1126/science.1105978; pmid: 15731445
3. J. F. Carbary, D. G. Mitchell, P. Brandt, E. C. Roelof, S. M. Krimigis, Statistical morphology of ENA emissions at Saturn. *J. Geophys. Res.* **113**, A05210 (2008). doi: 10.1029/2007JA012873
4. K. Dialynas et al., The extended Saturnian neutral cloud as revealed by global ENA simulations using Cassini/MIMI measurements. *J. Geophys. Res.* **118**, 3027–3041 (2013). doi: 10.1002/jgra.50295
5. P. Kollmann et al., A heavy ion and proton radiation belt inside of Jupiter's rings. *Geophys. Res. Lett.* **44**, 5259–5268 (2017). doi: 10.1002/2017GL073730
6. J. A. Van Allen, B. A. Randall, M. F. Thomsen, Sources and sinks of energetic electrons and protons in Saturn's magnetosphere. *J. Geophys. Res.* **85**, 5679–5694 (1980). doi: 10.1029/JA085iA11p05679
7. J. F. Cooper, "Innermost Van Allen Radiation Belt for High Energy Protons at Saturn" in *AAS/Division for Planetary Sciences Meeting Abstracts #40* (Bulletin of the American Astronomical Society, 2008), vol. 40, p. 460.
8. S. F. Singer, Trapped Albedo Theory of the Radiation Belt. *Phys. Rev. Lett.* **1**, 181–183 (1958). doi: 10.1103/PhysRevLett.1.181
9. J. B. Blake, H. H. Hilton, S. H. Margolis, On the injection of cosmic ray secondaries into the inner Saturnian magnetosphere: 1. Protons from the CRAND Process. *J. Geophys. Res.* **88**, 803–807 (1983). doi: 10.1029/JA088iA02p00803
10. R. S. Selesnick, M. K. Hudson, B. T. Kress, Direct observation of the CRAND proton radiation belt source. *J. Geophys. Res.* **118**, 7532–7537 (2013). doi: 10.1002/2013JA019338
11. E. Roussos et al., Long- and short-term variability of Saturn's ionic radiation belts. *J. Geophys. Res.* **116**, A02217 (2011). doi: 10.1029/2010JA015954
12. P. Kollmann, E. Roussos, A. Kotova, C. Paranicas, N. Krupp, The evolution of Saturn's radiation belts modulated by changes in radial diffusion. *New Astron.* **1**, 872–877 (2017). doi: 10.1038/s41550-017-0287-x
13. P. Kollmann et al., MeV proton flux predictions near Saturn's D ring. *J. Geophys. Res.* **120**, 8586–8602 (2015). doi: 10.1002/2015JA021621; pmid: 27812437
14. M. M. Hedman et al., Saturn's dynamic D ring. *Icarus* **188**, 89–107 (2007). doi: 10.1016/j.icarus.2006.11.017
15. M. M. Hedman et al., Three tenuous rings/arcs for three tiny moons. *Icarus* **199**, 378–386 (2009a). doi: 10.1016/j.icarus.2008.11.001
16. M. Hedman, J. Burt, J. Burns, M. Showalter, Non-circular features in Saturn's D ring: D68. *Icarus* **233**, 147–162 (2014). doi: 10.1016/j.icarus.2014.01.022
17. P. Kollmann, E. Roussos, C. Paranicas, N. Krupp, D. K. Haggerty, Processes forming and sustaining Saturn's proton radiation belts. *Icarus* **222**, 323–341 (2013). doi: 10.1016/j.icarus.2012.10.033
18. J. F. Cooper, Nuclear cascades in Saturn's rings: Cosmic ray albedo neutron decay and origins of trapped protons in the inner magnetosphere. *J. Geophys. Res.* **88**, 3945–3954 (1983). doi: 10.1029/JA088iA05p03945
19. X. Li et al., Measurement of electrons from albedo neutron decay and neutron density in near-Earth space. *Nature* **552**, 382–385 (2017). doi: 10.1038/nature24642; pmid: 29236693
20. A. Kotova, "Energetic particle tracking techniques and its application to the magnetosphere of Saturn," thesis, Universite Toulouse III – Paul Sabatier (2016).
21. I. G. Usoskin, K. Alanko-Huotari, G. A. Kovaltsov, K. Mursula, Heliospheric modulation of cosmic rays: Monthly reconstruction for 1951–2004. *J. Geophys. Res.* **110**, A12108 (2005). doi: 10.1029/2005JA011250
22. Materials and methods are available as supplementary materials.
23. T. P. Armstrong et al., Energetic ions trapped in Saturn's inner magnetosphere. *Planet. Space Sci.* **57**, 1723–1731 (2009). doi: 10.1016/j.pss.2009.03.008
24. E. Roussos et al., Sources, sinks and transport of energetic electrons near Saturn's main rings. *Geophys. Res. Lett.* 10.1029/2018GL078097 (2018).
25. J. A. VAN Allen, M. F. Thomsen, B. A. Randall, R. L. Rairden, C. L. Grosskreutz, Saturn's Magnetosphere, Rings, and Inner Satellites. *Science* **207**, 415–421 (1980). doi: 10.1126/science.207.4429.415; pmid: 17833551
26. S. M. Krimigis, T. P. Armstrong, Two-component proton spectra in the inner Saturnian magnetosphere. *Geophys. Res. Lett.* **9**, 1143–1146 (1982). doi: 10.1029/GL009i010p01143
27. R. E. Vogt et al., Energetic Charged Particles in Saturn's Magnetosphere: Voyager 2 Results. *Science* **215**, 577–582 (1982). doi: 10.1126/science.215.4532.577; pmid: 17771281
28. T. J. Birmingham, Charged particle motions in the distended magnetospheres of Jupiter and Saturn. *J. Geophys. Res.* **87**, 7421–7430 (1982). doi: 10.1029/JA087iA09p07421
29. M. F. Thomsen, J. A. Van Allen, Motion of trapped electrons and protons in Saturn's inner magnetosphere. *J. Geophys. Res.* **85**, 5831–5834 (1980). doi: 10.1029/JA085iA11p05831
30. J. F. Cooper, R. E. Johnson, P. Kollmann, E. Roussos, E. C. Sittler, "Plasma, neutral atmosphere, and energetic radiation environments of planetary rings," in *Planetary Ring Systems: Properties, Structure, and Evolution*, M. Tiscareno, C. Murray, Eds. (Cambridge Planetary Science Series, Cambridge Univ. Press, 2018), pp. 363–398.
31. R. S. Selesnick et al., Observations of the inner radiation belt: CRAND and trapped solar protons. *J. Geophys. Res.* **119**, 6541–6552 (2014). doi: 10.1002/2014JA020188
32. O. Adriani et al., Trapped proton fluxes at low Earth orbits measured by the PAMELA experiment. *Astrophys. J.* **799**, L4 (2015). doi: 10.1088/2041-8205/799/1/L4
33. S. J. Robbins, G. R. Stewart, M. C. Lewis, J. E. Colwell, M. Sremčević, Estimating the masses of Saturn's A and B rings from high-optical depth N-body simulations and stellar occultations. *Icarus* **206**, 431–445 (2010). doi: 10.1016/j.icarus.2009.09.012
34. J. E. Colwell et al., "The structure of Saturn's rings" in *Saturn from Cassini-Huygens*, M. K. Dougherty, L. W. Esposito, S. M. Krimigis, Eds. (Springer, 2009), pp. 375–412.
35. J. F. Cooper, J. H. Eraker, J. A. Simpson, The secondary radiation under Saturn's A-B-C rings produced by cosmic ray interactions. *J. Geophys. Res.* **90**, 3415–3427 (1985). doi: 10.1029/JA090iA04p03415
36. M. M. Hedman, P. D. Nicholson, The B-ring's surface mass density from hidden density waves: Less than meets the eye? *Icarus* **279**, 109–124 (2016). doi: 10.1016/j.icarus.2016.01.007
37. J. A. Van Allen, Absorption of energetic protons by Saturn's ring G. *J. Geophys. Res.* **88**, 6911–6918 (1983). doi: 10.1029/JA088iA09p06911
38. M. M. Hedman et al., The source of Saturn's G ring. *Science* **317**, 653–656 (2007). doi: 10.1126/science.1143964; pmid: 17673659
39. C. Paranicas, A. F. Cheng, A Model of Satellite Microsignatures for Saturn. *Icarus* **125**, 380–396 (1997). doi: 10.1006/icar.1996.5635
40. R. S. Selesnick, D. N. Baker, S. G. Kanekal, V. C. Hoxie, X. Li, Modeling the Proton Radiation Belt With Van Allen Probes Relativistic Electron-Proton Telescope Data. *J. Geophys. Res.* **123**, 685–697 (2018). doi: 10.1002/2017JA024661
41. R. S. Selesnick, M. D. Looper, R. A. Mewaldt, A theoretical model of the inner proton radiation belt. *Space Weather* **5**, S04003 (2007). doi: 10.1029/2006SW000275
42. P. Kollmann et al., Energetic particle phase space densities at Saturn: Cassini observations and interpretations. *J. Geophys. Res.* **116**, A05222 (2011). doi: 10.1029/2010JA016221
43. E. Roussos et al., Drift-resonant, relativistic electron acceleration at the outer planets: Insights from the response of Saturn's radiation belts to magnetospheric storms. *Icarus* **305**, 160–173 (2018). doi: 10.1016/j.icarus.2018.01.016
44. R. S. Selesnick, R. A. Mewaldt, Atmospheric production of radiation belt light isotopes. *J. Geophys. Res.* **101**, 19745–19757 (1996). doi: 10.1029/96JA01746
45. S. M. Krimigis, N. Sergis, D. G. Mitchell, D. C. Hamilton, N. Krupp, A dynamic, rotating ring current around Saturn. *Nature* **450**, 1050–1053 (2007). doi: 10.1038/nature06425; pmid: 18075586
46. J. E. P. Connerney et al., A New Model of Jupiter's Magnetic Field From Juno's First Nine Orbits. *Geophys. Res. Lett.* **45**, 2590–2596 (2018). doi: 10.1002/2018GL077312

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/362/6410/eaat1962/suppl/DC1
Materials and Methods
Figs. S1 to S7
Table S1
References (47–56)

RESEARCH ARTICLE SUMMARY

GAS GIANT PLANETS

The low-frequency source of Saturn's kilometric radiation

L. Lamy*, P. Zarka, B. Cecconi, R. Prangé, W. S. Kurth, G. Hospodarsky, A. Persoon, M. Morooka, J.-E. Wahlund, G. J. Hunt

INTRODUCTION: Planetary auroral radio emissions are powerful nonthermal cyclotron radiation produced by magnetized planets. Their remote observation provides information on planetary auroral processes and magnetospheric dynamics. Understanding how they are generated requires in situ measurements from within their source region. During its early 2008 high-inclination orbits, the Cassini spacecraft unexpectedly sampled two local sources of Saturn's kilometric radiation (SKR) at 10-kHz frequencies, corresponding to the low-frequency (LF) portion of its 1- to 1000-kHz typical spectrum (hence only observable from space). These provided insights into the underlying physical excitation mechanism. The combined analysis of radio, plasma, and magnetic field in situ measurements demonstrated that the cyclotron maser instability (CMI), which generates auroral kilometric radiation (AKR) at Earth, is a universal generation mechanism able to operate in widely different planetary plasma environments. The CMI requires accelerated (out-of-equilibrium) electrons and low-density magnetized plasma, such as auroral regions where the ratio of the electron plasma frequency f_{pe} to the electron cyclotron frequency f_{ce} is much lower than unity.

RATIONALE: Intensifications of the SKR spectrum in general, and of its LF part in particular, have been widely used as a diagnostic of Saturn's large-scale magnetospheric dynamics, such as the SKR rotational modulation or major auroral storms driven by the solar wind. However, the limited set of events encountered in 2008 did not provide a comprehensive picture of the source of SKR LF emissions and of the conditions in which the CMI can trigger them.

During the 20 ring-grazing high-inclination orbits and the preceding 7 orbits, which together spanned late 2016 to early 2017, the Cassini spacecraft repeatedly sampled the top of the SKR emission region, at distances of a few planetary radii (R_S), corresponding to the lowest-frequency part of the SKR spectrum. SKR emission frequency is close to f_{ce}, itself proportional to the magnetic field amplitude, and so decreases with increasing distance from the planet. We conducted a survey of these orbits to extract the average properties of the radio sources encountered and assess the ambient magnetospheric plasma parameters that control them.

RESULTS: Throughout this set of trajectories, we were able to identify only three SKR sources. They covered the 10- to 20-kHz range (3.5 to 4.5 R_S distances from the planet's center) and were solely found on the northern dawn-side sector. The source regions hosted narrow-banded emission, propagating in the extraordinary wave mode, and radiated quasi-perpendicularly to the magnetic field lines. Their emission frequency, measured in situ, was fully consistent with the CMI mechanism driven by 6- to 12-keV electron beams with shell-type velocity distribution functions, as for AKR at Earth.

The SKR sources were embedded within larger regions of upward currents, themselves coincident with the ultraviolet (UV) auroral oval, which was observed simultaneously with the Hubble Space Telescope. However, unlike the terrestrial case, the spacecraft exited the radio source region before exiting the main oval, itself strictly coincident with the upward current layer, when the ratio f_{pe}/f_{ce} exceeded the typical CMI threshold of 0.1. This occurred at times of sudden local increases of the magnetospheric electron density.

CONCLUSION: The generation conditions of SKR LF emission appear to be strongly time-variable. The SKR spectrum additionally displays a significant local time dependence, with the lowest frequencies (the highest altitudes) reached for dawn-side radio sources. The characteristics of CMI-unstable electrons at 3.5 to 4.5 R_S imply that downward electron acceleration took place at farther distances along the auroral magnetic field lines and brings new constraints to particle acceleration models. Finally, the LF SKR component is mainly controlled by local plasma conditions, namely time-variable magnetospheric electron densities, which can quench the CMI mechanism even if accelerated electrons are present. This explains why the magnetic field lines hosting SKR LF sources can map to a restricted portion of the UV auroral oval and associated upward current region. ∎

Schematic view of Saturn's auroral emissions.
Auroral radio emissions observed above the atmosphere were remotely mapped from Cassini (the most intense emissions in red) when the spacecraft encountered a low-frequency radio source. The magnetic field lines hosting it, sampled in situ by the probe, map to the (atmospheric) UV auroral oval, as observed simultaneously by the Hubble Space Telescope.
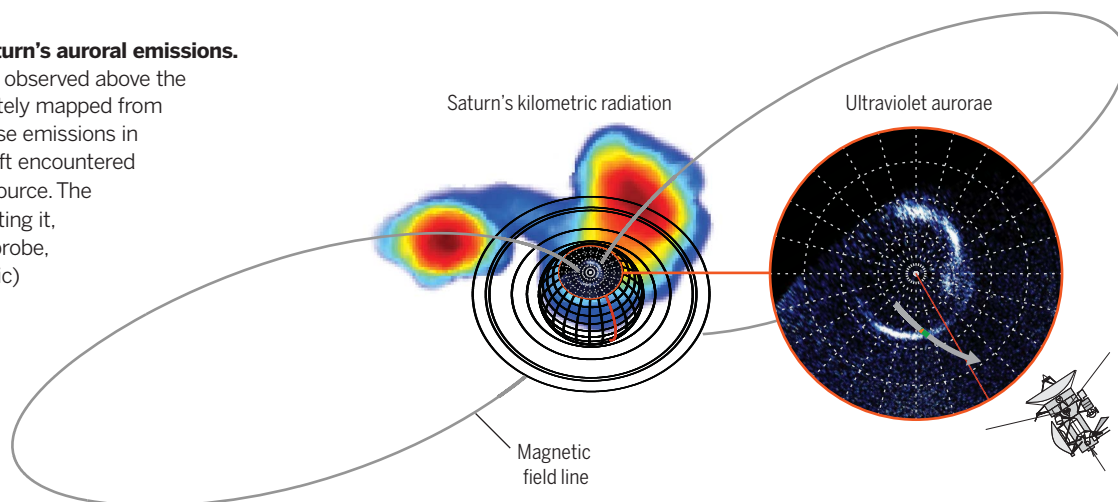
Saturn's kilometric radiation

Ultraviolet aurorae

Magnetic field line

RESEARCH ARTICLE

GAS GIANT PLANETS

The low-frequency source of Saturn's kilometric radiation

L. Lamy[1]*, P. Zarka[1], B. Cecconi[1], R. Prangé[1], W. S. Kurth[2], G. Hospodarsky[2], A. Persoon[2], M. Morooka[3], J.-E. Wahlund[3], G. J. Hunt[4]

Understanding how auroral radio emissions are produced by magnetized bodies requires in situ measurements within their source region. Saturn's kilometric radiation (SKR) has been widely used as a remote proxy of Saturn's magnetosphere. We present wave and plasma measurements from the Cassini spacecraft during its ring-grazing high-inclination orbits, which passed three times through the high-altitude SKR emission region. Northern dawn-side, narrow-banded radio sources were encountered at frequencies of 10 to 20 kilohertz, within regions of upward currents mapping to the ultraviolet auroral oval. The kilometric waves were produced on the extraordinary mode by the cyclotron maser instability from 6– to 12–kilo–electron volt electron beams and radiated quasi-perpendicularly to the auroral magnetic field lines. The SKR low-frequency sources appear to be strongly controlled by time-variable magnetospheric electron densities.

S aturn is a powerful planetary radio emitter, dominated in intensity by Saturn's kilometric radiation (SKR). This nonthermal auroral emission, equivalent to auroral kilometric radiation (AKR) at Earth, is produced at frequencies f close to the local electron cyclotron frequency f_{ce}, itself proportional to the magnetic field amplitude and thus decreasing with increasing distances from the planet. The SKR spectrum ranges from a few kHz to 1 MHz (thus only observable from space), which corresponds to radio sources distributed along auroral field lines from slightly above the ionosphere up to several planetary radii R_S (1 R_S = 60,268 km is Saturn's radius). Remote observations of SKR have long been used as a proxy of kronian auroral processes and magnetospheric dynamics (*1*).

During its early 2008 high-inclination passes, the Cassini spacecraft unexpectedly encountered two SKR low-frequency (LF) source regions. The combined analysis of radio, plasma, and magnetic field in situ measurements showed that the cyclotron maser instability (CMI), which generates AKR at Earth, also produces SKR at Saturn, suggesting a universal process able to operate in widely different planetary magnetospheres. The CMI is a wave-electron resonant interaction that requires (i) accelerated (out-of-equilibrium) weakly relativistic electrons, whose velocity distribution function displays a positive gradient along the velocity direction perpendicular to the magnetic field, and (ii) an electron plasma frequency f_{pe} much lower than f_{ce} (*2*). When these two conditions are fulfilled, radio emission is amplified, mainly on the right-hand extraordinary (R-X) wave propagation mode, at the frequency $f_{CMI} = f_{ce}/\Gamma + k_\parallel v_\parallel/2\pi$ where **k** is the wave vector, **v** the electron velocity, the $\parallel$ subscript refers to the direction parallel to the magnetic field and $\Gamma = 1/\sqrt{1 - v^2/c^2}$ is the Lorentz factor (*3*). Waves are thus amplified at the local relativistic electron cyclotron frequency, Doppler-shifted by the parallel propagation. To first order, this expression approximates to $f_{CMI} \sim f_{ce}$.

A first SKR source was identified on day 2008-291 (17 October) at 10 kHz (corresponding to 5 R_S distance from Saturn's center), at the unexpected Local Time (LT) of 01:00 and along magnetic field lines with −80° invariant latitude. Source traversal occurred during a global SKR intensification extending to low frequencies with wave electric fields of a few mV m^{-1}. It coincided with strong upward-directed field-aligned currents, indicating down-going electron beams, triggered by a major solar wind–driven compression of the magnetosphere (*4*–*6*). The R-X mode source was identified from measurements of its low-frequency cutoff f_{SKR} at about 2% below f_{ce}. Radio waves displayed strong linear polarization at the source from which they were radiated quasi-perpendicularly to the local magnetic field. In this case, f_{SKR} can be expressed as $f_{CMI} = f_{ce}\sqrt{1 - v^2/c^2}$ and direct measurements of f_{SKR} yielded CMI-resonant electrons of 6 to 9 keV, in agreement with the hot (accelerated) electrons simultaneously measured in situ. Their shell-like three-dimensional (3D) distribution function (referred to as horseshoe in 2D) was

also fully consistent with CMI-driven perpendicular emission. Sampling down-going electrons with shell-like distribution implies that they had been accelerated toward the planet in regions located farther than the radio source along the flux tube (*4*, *7*–*9*). A second direct confirmation of the CMI mechanism was brought by the agreement between the wave properties observed within the source region and the simulated CMI wave growth rate and emission angle driven by the electron distribution function, measured simultaneously (*9*). As at Earth, the auroral plasma surrounding the emission region was hot, with a marginal cold component, and tenuous with $f_{pe}/f_{ce} \sim 0.1$ (*10*). Conversely, terrestrial-like plasma cavities were not observed within the radio source at the 2-s resolution of Cassini's electron measurements. This difference, together with the presence of SKR sources at larger distances from the planet than AKR sources at Earth, was attributed to the faster rotation of Saturn's magnetosphere. This means that CMI condition (ii) is generally fulfilled at high latitudes without needing small-scale cavities.

A second SKR source was likely traversed on day 2008-073 (13 March) at 5 kHz (i.e., 6 R_S distance) (*11*). The LF emission tangentially approached the f_{ce} curve on the corresponding dynamic spectrum (time-frequency map of wave flux density) at a more usual location, near 07:30 LT, close to the SKR peak intensity region (*12*), and along magnetic field lines with −77° invariant latitude. The f_{pe}/f_{ce} ratio was as low as 0.05, again fulfilling CMI condition (ii). The low-frequency cutoff f_{SKR} was not accurately measured, but 7-keV electrons with a shell-like distribution function yielded CMI wave growth rates compatible with the observed SKR R-X mode flux densities (*13*).

This limited set of events did not provide a comprehensive view of SKR LF sources and of the conditions under which the CMI can trigger them. Intensifications of SKR spectrum in general, and of its LF component in particular, have nevertheless long been used as a sensitive diagnostic of large-scale magnetospheric dynamics, including the SKR dual rotational modulation or auroral storms induced (or not) by the solar wind (*14*–*25*). The final high-inclination phase of the Cassini mission carried the spacecraft to high latitudes at low altitudes, providing the opportunity to statistically investigate SKR sources in situ. Those orbits began with the ring-grazing orbital sequence, during which the spacecraft repeatedly sampled the top of both northern and southern SKR emission regions, at distances corresponding to the SKR lowest frequencies. We present a survey of these orbits, extract the average properties of the radio sources encountered, and discuss the magnetospheric parameters that control them.

Survey of ring-grazing orbits

Cassini's 20 ring-grazing orbits (numbered 251 to 270), whose periapsis was outside the F-ring, were executed between days 2016-335 (30 November) and 2017-112 (22 April). They

¹Laboratoire d'Etudes Spatiales et d'Instrumentation en Astrophysique, Observatoire de Paris, Université Paris Sciences et Lettres, Centre National de la Recherche Scientifique, Sorbonne Université, Université Paris Diderot, Sorbonne Paris Cité, 5 Place Jules Janssen, 92195 Meudon, France. ²Department of Physics and Astronomy, University of Iowa, Iowa City, IA 52242, USA. ³Swedish Institute of Space Physics, Box 537, SE-751 21 Uppsala, Sweden. ⁴Blackett Laboratory, Imperial College London, London SW7 2BW, UK.
*Corresponding author. Email: laurent.lamy@obspm.fr

passed through auroral field lines expected to host SKR sources up to 30 kHz (40 kHz) in the dawn-side northern (dusk-side southern) hemisphere (fig. S1). They were preceded by seven orbits (numbered 244 to 250) starting on day 2016-270 (26 September) that brought the spacecraft to high latitudes and at distances close enough to the planet to probe SKR sources at a few kHz. We conducted a survey of kilometric sources encountered during these 27 orbits using combined observations of the Radio and Plasma Wave Science (RPWS) experiment (26) and the magnetometer (MAG) (27).

After the Cassini Plasma Spectrometer (CAPS) (28) ceased operating in 2012, it was no longer possible to measure in situ low-energy electron distribution functions. This prevented the identification of SKR sources from the direct analysis of simultaneous radio and electron measurements based on growth rate calculation. Instead, on the basis of the reference event of day 2008-291, SKR sources were identified solely from RPWS-HFR (high-frequency receiver) and MAG data whenever (j) the emission LF cutoff f_{SKR} passed strictly below f_{ce} and (jj) persisted for at least two consecutive RPWS-HFR measurements, obtained at a 16-s temporal resolution. Criterion (j) rests upon the assumption of perpendicular emission (examined below). Criterion (jj) conservatively avoids false detections induced by spiky signals (see examples in fig. S2). The method used to determine f_{SKR} is reminded in (2). Our survey revealed only three events that unambiguously fulfilled both criteria. They all occurred in the northern hemisphere on days 2016-339 (4 December), 2016-346 (11 December), and 2017-066 (7 March), hereafter referred to as events S_1, S_2, and S_3. No source was found in the southern hemisphere. We rejected several candidate northern SKR sources that fulfilled criterion (j) but not (jj). These may be either nearby sources (close to but not crossed by Cassini), crossings of small-sized sources lasting less than 16 s (~200 to 300 km for typical spacecraft velocity), and/or emission produced at $f_{CMI} > f_{ce}$. The latter case is possible for oblique emission, where the $k_\parallel v_\parallel$ term in the expression of f_{CMI} becomes non-negligible (8, 9).

Events S_1, S_2, and S_3 were sampled during intervals of relatively quiet solar wind activity (29). S_1 and S_2 additionally occurred close to the predicted timing of periodic northern SKR bursts, whereas event S_3 was instead in anti-phase with it (1, 22). The first two events were thus consistent with standard auroral activity, whereas S_3 indicated a more unusual case, possibly driven by internal dynamics. The detailed wave properties and characteristics of the three identified sources are listed in Table 1. Hereafter, we focus on events S_3 and S_1 as illustrative examples. A polar projection of the magnetic footprint of these sources is provided in fig. S3, which shows that they are additionally colocated with the average northern ultraviolet (UV) aurorae, as observed by the Hubble Space Telescope (HST) throughout 2017 (29). Event S_3 displayed the most intense radio emission and benefited from

Table 1. Properties of the three R-X mode SKR low-frequency sources identified during the Cassini ring-grazing orbits. Where provided, uncertainties are one standard deviation.

SKR source	S_1	S_2	S_3
Time range (UT)	2016-339 08:42:50–08:59:54 (orbit 251)	2016-346 13:33:26–13:34:30 (orbit 252)	2017-066 15:11:48–15:17:25 (orbit 264)
Cassini ephemeris:			
Latitude (°)	+61.9–60.9	+59.6–59.5	+54.0–53.1
LT (hours)	9.06–9.41	9.72–9.74	10.32–10.43
Distance to Saturn's center (R_S)	4.47–4.31	4.15–4.14	3.63–3.58
Altitude along the flux tube (R_S) from the 1-bar atmospheric level	3.65–3.49	3.33–3.32	2.84–2.79
f_{ce} (kHz)	13–14	15	22–23
Δf (kHz)	0.1–0.4	<1.5	<2
Size range (km)	240–3400 (9 subsources)	1300 (1 source)	6050 (1 source)
Mean f_{SKR}/f_{ce} ratio (range)	0.976 ± 0.006 (0.965–0.998)	0.975 ± 0.008 (0.968–0.987)	0.989 ± 0.004 (0.978–0.995)
Mean electron energy (keV) (range)	11 ± 3 (1–17)	12 ± 4 (7–16)	6 ± 3 (3–11)
Median Poynting flux (W/m²) (range)	$5.9\ 10^{-13}$ ($2.6\ 10^{-14}$–$2.7\ 10^{-11}$)	$1.2\ 10^{-13}$ ($9.1\ 10^{-14}$–$1.6\ 10^{-12}$)	$7.3\ 10^{-12}$ ($2.0\ 10^{-13}$–$2.3\ 10^{-8}$)
Median electric field amplitude (mV/m) (range)	0.033 ± 0.032 (0.004–0.142)	0.014 ± 0.011 (0.008–0.035)	0.292 ± 0.884 (0.012–4.211)
Beaming angle (°) $\theta = (\mathbf{k}, \mathbf{B})$	93 ± 9	86 ± 16	91 ± 16
Upward current density (nA/m²)	157	171	105

coordinated HST observations, whereas event S_1 was sampled by RPWS-WBR (wideband receiver) waveform measurements.

Example SKR source crossings

Radio, plasma, and magnetic field measurements during events S_3 and S_1 are shown in Figs. 1 and 2, respectively (S_2 is shown in fig. S2). SKR is the patchy bright emission observed above 10 kHz in Figs. 1A and 2A and fig. S2A (dynamic spectrum of the wave Poynting flux S), with dominant right-handed (RH) circular polarization in Figs. 1B and 2B and fig. S2B (dynamic spectrum of the normalized degree of circular polarization V) indicative of northern R-X mode. Some weaker left-handed (LH) polarized SKR, indicative of left-hand ordinary (L-O) mode, was additionally observed at the low-frequency edge of the R-X mode SKR toward the end of the intervals plotted in Figs. 1 and 2 and fig. S2, preceding the periapsis. The weakly polarized emission visible below 10 kHz after 15:30 UT (Universal Time) on Fig. 1, A and B, is identified as auroral hiss, a whistler-mode emission. Narrow-banded long-duration and drifting short-duration emissions observed below 35 kHz are not investigated in our analysis. Localized SKR R-X mode LF emission is tangent to the f_{ce} curve near the middle of each interval. Condition $f_{SKR} < f_{ce}$ (Figs. 1C and 2C and fig. S2C) is then used to define the overall extent of the source region. During event

S_3 (Fig. 1, A to C), a single SKR source was continuously observed from 15:11:48 to 15:17:25 UT, corresponding to a 6050-km distance along the spacecraft trajectory. During event S_1 (Fig. 2, A to C), multiple SKR sources were observed between 08:42:50 and 08:59:54 UT.

Figure 3 shows restricted portions of Fig. 2A at higher time-frequency resolution obtained from RPWS-WBR waveform measurements (2). Faint, localized, remote SKR emissions observed well above f_{ce} (Fig. 3A) display recurrent drifting substructures with scarce positive and prominent negative rates of ~1 kHz/min, consistent with previous SKR waveform observations (30). An enlargement of the 08:00-09:10 UT interval (Fig. 3B) reveals the time-frequency structure of the radio source. Whereas Fig. 1A suggested a single emission pattern with a bandwidth of a few kHz lasting for nearly 2 hours, Fig. 3B shows that it splits into several subcomponents of variable durations (down to a few min) and bandwidths. Those with frequencies slightly, but clearly, above f_{ce} likely correspond to sources on nearby flux tubes, while only the component strictly tangent to and below the f_{ce} curve unambiguously indicates a local source. The overplotted f_{SKR} profile (as determined in Fig. 2C) provides, despite its lower temporal resolution, a one-to-one correspondence between the brief episodes of $f_{SKR} < f_{ce}$ and the observed SKR bursts (including the central event between 08:45 and 08:50 UT and

the shorter bursts beside it). This agreement validates the method employed to determine f_{SKR}. For this specific event, we therefore considered nonconsecutive RPWS-HFR episodes with $f_{SKR} < f_{ce}$ as isolated subsources as well. The overall source region of event S_1 thus decomposes into nine SKR subsources. Their horizontal size along the spacecraft trajectory, and across the encountered flux tubes, ranged from 240 to 3400 km. Their ~100- to 400-Hz bandwidth Δf characterizes very narrow-banded emissions with $\Delta f/f_{ce}$ 0.008 to 0.03, and thus corresponds to radiating structures that extended vertically 700 to 2800 km along the magnetic flux tubes.

Perpendicular emission and energy of source electrons

Figures 1 and 2C and fig. S2C (right) show the kinetic energy of CMI-unstable electrons derived

from the measurements of $f_{SKR} < f_{ce}$, assuming perpendicular emission at the source and thus $f_{SKR} = f_{ce}\sqrt{1 - v^2/c^2}$, as for the southern event of 2008-291. To check this assumption, we analyzed the wave parameters derived from the goniopolarimetric (direction-finding) analysis of RPWS-HFR three-antenna measurements (2). These consist of the wave Stokes parameters S, Q, U, and V (31) and the $\boldsymbol{k}$-vector direction. From the latter, we also retrieved the 3D locus of the radio source, from the intersection of the $\boldsymbol{k}$ line-of-sight with a model iso-$f = f_{ce}$ surface (assuming straight-line propagation, which is reasonable at high latitude), and its beaming angle $\theta = (\boldsymbol{k}, \boldsymbol{B})$, where $\boldsymbol{B}$ is the magnetic field vector at the source (4, 8, 12, 32).

Figure 4 displays the variation of the beaming angle $\theta(f)$ of events S_3 and S_1 for the time intervals of Figs. 1 and 2 using a standard data selec-

tion (2) for RH (LH) polarized waves, and thus R-X (L-O) mode. In agreement with previous studies of the SKR beaming angle (8, 23, 32, 33), $\theta(f)$ of R-X mode emission smoothly rises from 60 ± 25° at 800 kHz to 85°±15° at 30 kHz, where the uncertainty on θ is dominated by the scatter of data points. Figure S4E shows an alternative representation of $\theta(f)$ in the time-frequency plane. This systematic change from perpendicular to oblique emission with increasing frequency is most likely due to refraction of R-X mode waves on the nearby iso-f_X surface, where $f_X \sim f_{ce}$ is the R-X mode cutoff frequency (8), as proposed for auroral radio emissions of Earth and Jupiter (34, 35). In this case, the apparent R-X mode beaming angle should vary as a function of the $\boldsymbol{k}$-vector direction projected in a plane perpendicular to $\boldsymbol{B}$, with the strongest refraction expected for wave propagation toward the

Fig. 1. Cassini radio, plasma, and magnetic measurements acquired on day 2017-066 between 13:00 and 17:00 UT (event S₃) in the northern dawn-side sector. (**A** and **B**) Dynamic spectra of wave Poynting flux S and normalized degree of circular polarization V, respectively, derived from two- and three-antenna RPWS-HFR 16-s measurements over 3.5 to 1500 kHz (2). $V = -1$ (+1) refer to RH (LH) polarization, and thus to R-X (L-O) mode emission from the northern hemisphere. The white dashed lines indicate the electron cyclotron frequency f_{ce} as derived from MAG data. R-X mode SKR around 20 kHz comes close to f_{ce} after 15:00 UT. (**C**) Ratio of the SKR low-frequency cutoff f_{SKR} to f_{ce} (2). Values $f_{SKR}/f_{ce} < 1$ define one SKR source crossing between 15:11 and 15:18 (orange-shaded region on (C) and (D), extended by vertical dotted-dashed lines in the other panels). The right-handed gray scale provides the kinetic electron energy inferred from $f_{SKR} = f_{ce}\sqrt{1 - v^2/c^2}$, assuming perpendicular emission. (**D**) Characteristic frequencies comparing f_{SKR} (solid black line), f_{ce}, and $f_{ce}/10$ (dashed black lines) and the electron plasma frequency f_{pe} as directly derived from RPWS Langmuir probe 32-s measurements (gray profile) together with a lower limit $f_{pe\ min}$ estimated from the auroral hiss upper frequency cutoff measured with RPWS-MFR 64-s measurements (2) (red profile). (**E**) Magnetic azimuthal component $B\varphi$, as from 1-min averaged MAG measurements in spherical coordinates. The green-shaded area marks a region of positive slope, indicative of an upward-directed current region. The SKR source region starts at the beginning of the upward current region and ends before Cassini exits it, when f_{pe} suddenly rises above $f_{ce}/10$.
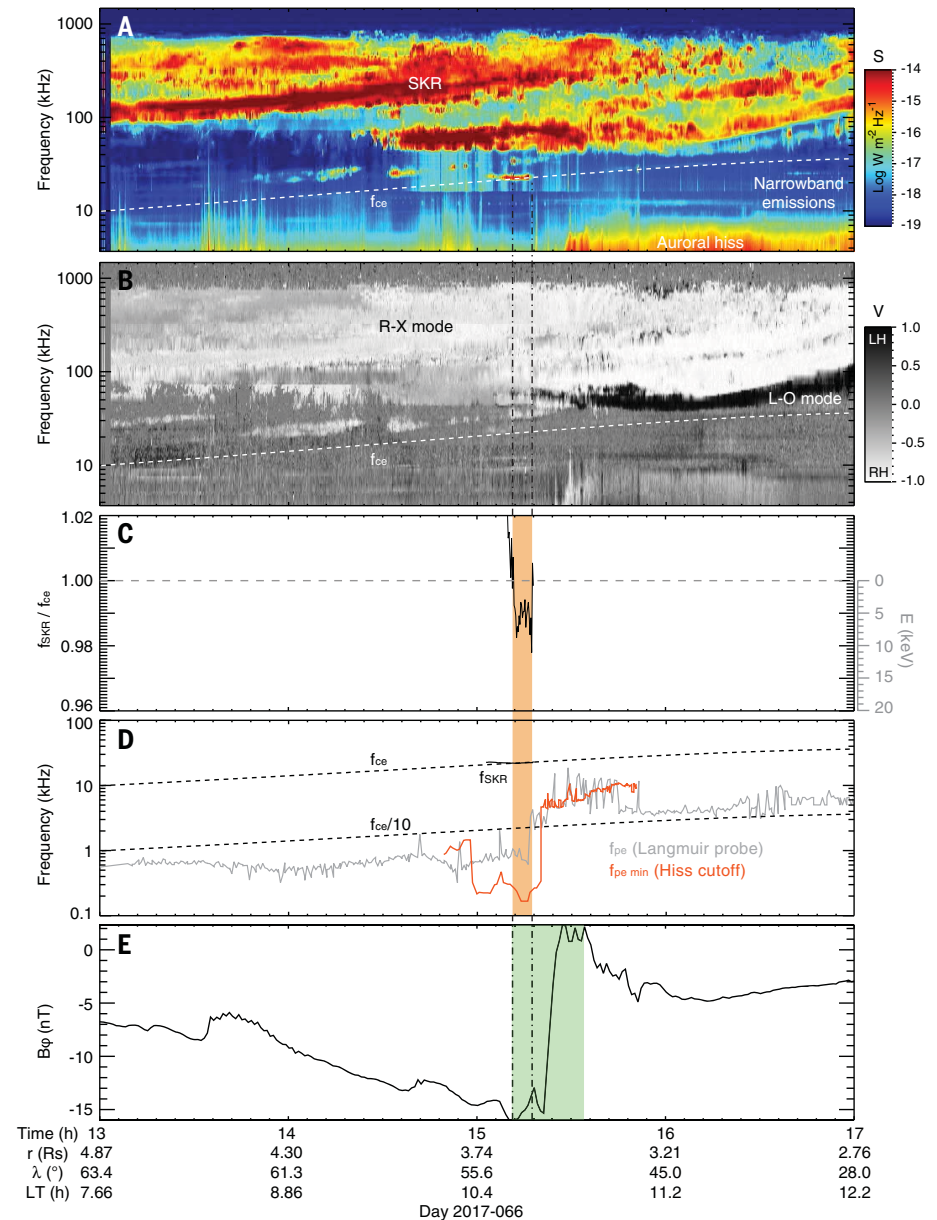
equator. L-O mode SKR was also observed, over an extended spectral range of 30 to 100 kHz, with θ(f) remaining confined above 75° and with a lower scattering than the R-X mode. The L-O mode cutoff frequency f_O is f_{pe}, which was in this case much lower than f_{ce} (Figs. 1 and 2D), so no similar refraction on the iso-f_O surface is expected. This interpretation implies that θ(f) measures the apparent (real) beaming angle of SKR R-X (L-O) mode affected (unaffected) by refraction close to the source. Although they differ at high frequencies, both remote observations of R-X and L-O mode θ(f) indicate quasi-perpendicular emission at 30 kHz, toward frequencies where the local radio sources have been sampled.

A second method provides a direct measurement of the wave emission angle within the source from its polarization state. For perpendicular cyclotron emission, the polarization ellipse axial ratio $AR = (T - L)/V$, where $T = \sqrt{Q^2 + U^2 + V^2}$ and $L = \sqrt{Q^2 + U^2}$ are the total and linear polarization degrees (31), can be approximated to first order as cos θ (36). Figures 1 and 2B and fig. S4, A to D, show that the emission is strongly linearly polarized at the source, as observed during the southern event of 2008-291. Measurement of the mean AR in each SKR source of the three events S_1, S_2, and S_3, restricting to the two HFR frequencies encompassing f_{ce} with standard data selection (identical to that used in Fig. 5C, discussed below), gives $AR = -0.02 \pm 0.26$, 0.06 ± 0.28, and -0.05 ± 0.15, which results in θ = 91 ± 16°, 86 ± 17°, and 93 ± 9°, respectively, all consistent with perpendicular emission.

The measured values of f_{SKR} can therefore be safely inverted to derive kinetic energies of CMI-unstable electrons (Figs. 1C and 2C, fig. S2C, and Table 1). On average, f_{SKR} reached a level of −2% from f_{ce} with a minimum value of −3.5%. The corresponding electron energies spanned the 1- to 17-keV range, with mean values between 6 and 12 keV depending on the event and a standard deviation of 3 to 4 keV, larger than the 2-keV uncertainty arising from the f_{SKR} measurements. No correlation is found between the obtained electron energies and the observed wave intensities, measured between a few 0.01 and a few mV m⁻¹.

Auroral plasma conditions

We now investigate the auroral plasma conditions to assess the validity of CMI condition (ii). Figures 1D and 2D compare the characteristic frequencies f_{SKR}, f_{ce}, and f_{pe}. f_{pe} was derived from Langmuir probe (LP) 32-s measurements and compared to the upper frequency cutoff of auroral hiss derived independently from the RPWS-MFR (medium frequency receiver) 64-s measurements. The LP measures the total electrical current carried by the ambient plasma from voltage sweeps, from which the local electron density N and thus f_{pe} are derived (2). The voltage sweep resolution leads to a typical uncertainty on f_{pe} of less than a factor of 2. Auroral hiss is radiated upward from below the spacecraft and propagating beneath f_{pe}, so that its upper fre-
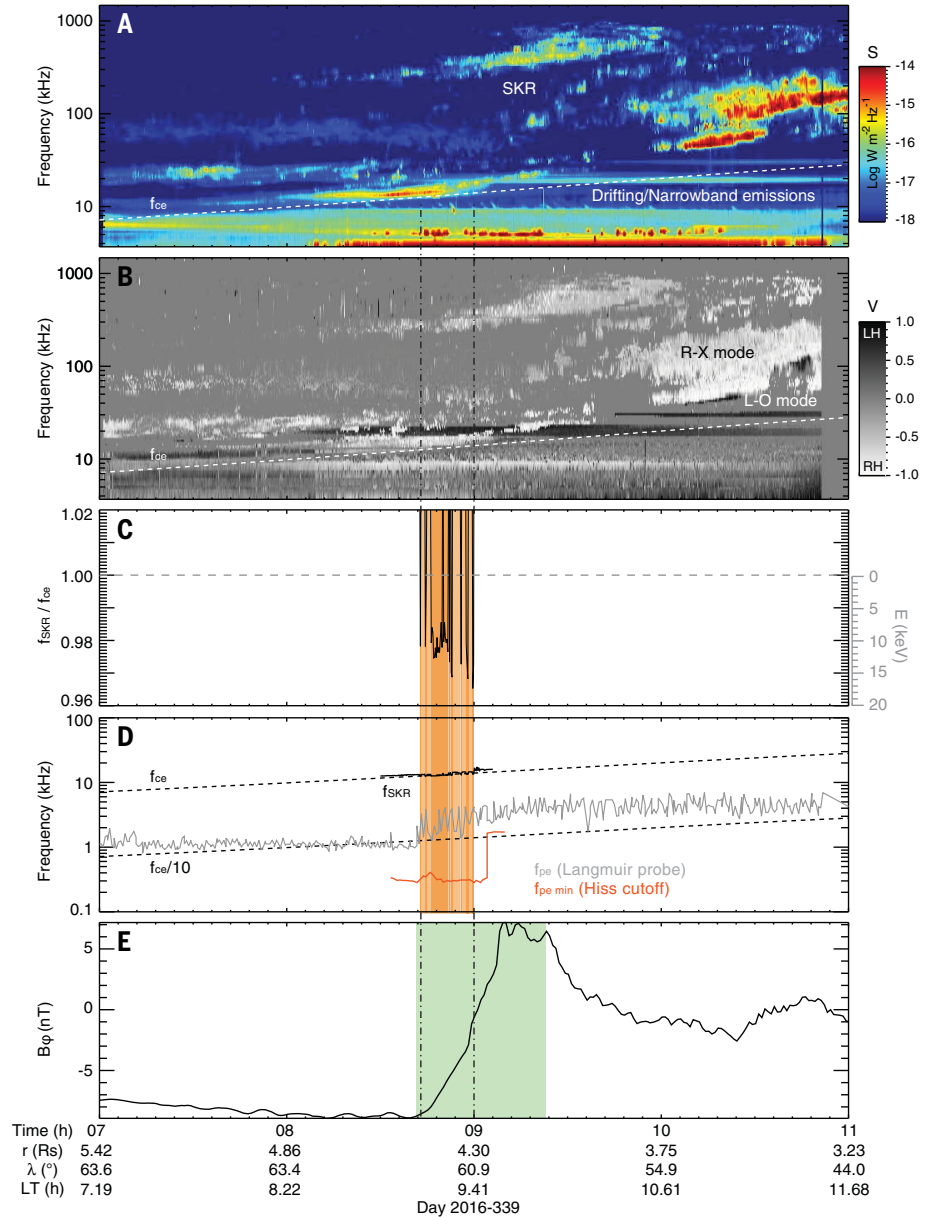
Fig. 2. As for Fig. 1, but for measurements acquired on day 2016-339 between 07:00 and 11:00 UT (event S₁) in the northern dawn-side sector. (**A** to **E**) Nine SKR sources near 13 kHz were encountered successively between 08:43 and 09:00. The overall interval is delimited by a light orange-shaded region, while the individual SKR sources are indicated by dark orange-shaded areas.

quency cutoff, when marking a sharp transition, is a proxy of f_{pe} when $f_{pe} < f_{ce}$. The observed hiss emissions were generally damped and exhibited a smoothed cutoff, so we interpret the fitted frequencies as a lower limit on f_{pe}, labeled $f_{pe\ min}$. In Figs. 1D and 2D, the $f_{pe\ min}$ profiles were close to, and generally below, the LP-derived f_{pe} profiles. For event S_1 (Fig. 1D), f_{pe}/f_{ce} was about 0.03 across the whole source region, which ended when the value of f_{pe} suddenly rose by a factor of 4 at 15:18 UT (and $f_{pe\ min}$ by a factor of 13 at 15:20 UT), exceeding the $f_{ce}/10$ threshold for CMI condition (ii). For event S_3 (Fig. 2D), a similar density gradient was observed in both data-

sets. The hiss cutoff frequency $f_{pe\ min}$ rose by a factor of 5 at 09:04 UT, again shortly after exiting the SKR source region at 09:00 UT. The LP-derived f_{pe}, determined at a higher time resolution, reveals more complex variability: Although f_{pe} rose by a factor of 4 at 08:42, it then showed large-scale fluctuations of a factor of ~3 (~2) up to (after) 09:00 UT. The SKR subsources (sampled at a better cadence) match the local minima of f_{pe}, suggesting that small-sized radio sources were colocated with local density cavities. Similar results were observed during event S_2 (fig S2D), despite lower time-resolution measurements of f_{pe}.
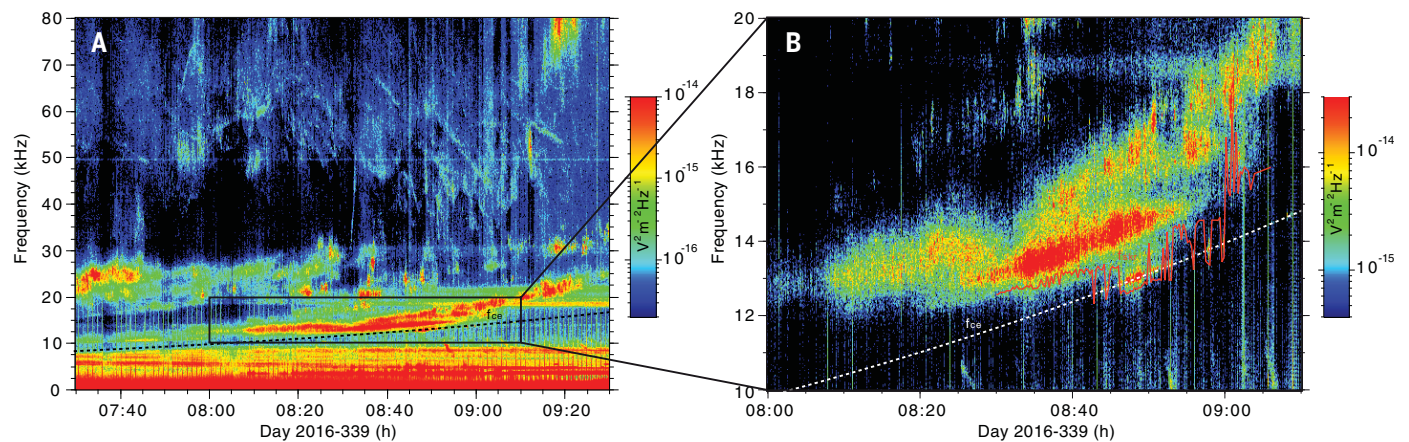
Fig. 3. High-resolution dynamic spectra of event S_1 reconstructed from Cassini radio electric waveform measurements taken on 2016-339. The spectra were acquired with the RPWS-WBR receiver over the 0- to 80-kHz range (2). (**A**) Reconstructed dynamic spectrum between 07:30 and 09:30 UT. The dashed line indicates f_{ce}. (**B**) Enlargement of the encountered local SKR source. The solid red line indicates f_{SKR}, as derived from RPWS-HFR measurements in Fig. 2D. It is in agreement with high-resolution WBR measurements of SKR emission close to and below f_{ce}.
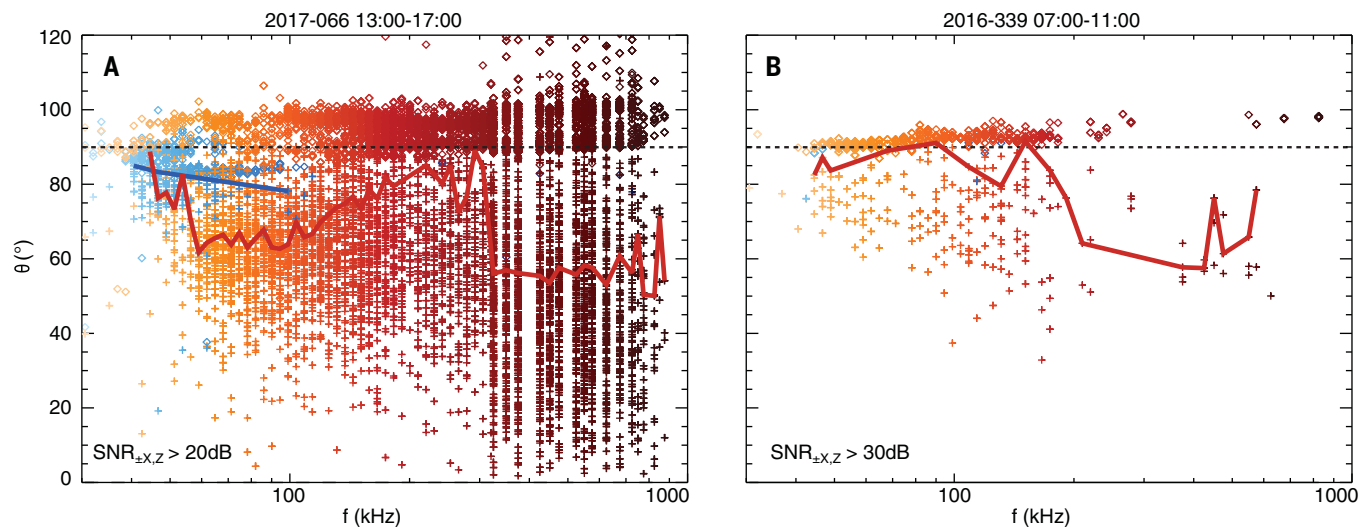
Fig. 4. SKR beaming angle at the source as a function of frequency. (**A**) Time interval from 13:00 to 17:00 UT on 2017-066 (event S_3) encompassing the SKR source crossed around 15:15 UT. The beaming angle is defined as $\theta = (\mathbf{k}, \mathbf{B})$, where $\mathbf{k}$ is the wave vector determined remotely from RPWS-HFR three-antenna measurements, and $\mathbf{B}$ the magnetic field vector at the source (2). A standard data selection was applied with $T > 0.85$, $|z_r| < 0.05$, and signal-to-noise ratio $SNR_{\pm X,Z} > 20$ dB. The color scale refers to the wave frequency, with red (blue) symbols corresponding to RH (LH) polarized waves and thus to R-X (L-O) mode. Crosses (diamonds) indicate $\mathbf{k}$ directions that intercept (do not intercept) their iso-$f = f_{ce}$ surface, assuming straight-line propagation. Red and blue lines indicate the median value of θ, derived at each frequency for which the number of crosses exceeds the number of diamonds (7). The dotted line marks emission perpendicular to the local magnetic field vector. (**B**) Time interval from 07:00 to 11:00 UT on 2016-339 (event S_1) encompassing the multiple SKR sources crossed around 08:50 UT. The data selection was identical to that of (A) except that $SNR_{\pm X,Z} > 30$ dB to better remove faint non-SKR narrow-banded emissions observed up to 40 kHz (2). Both panels show that θ increases from $60 \pm 25°$ at 800 kHz to $85 \pm 15°$ at 30 kHz, consistent with quasi-perpendicular emission at the source.

Figures 1E and 2E and fig. S2E show the azimuthal component of the magnetic field measured from 1-min averaged MAG data. They indicate upward currents in intervals dominated by a positive slope. Each SKR source region started at (but ended before) the beginning (end) of the upward current region. The upward current density was estimated to be 106, 171, and 157 nA m^{-2} for events S_1, S_2, and S_3, respectively, well above the average of 80 nA m^{-2} derived from all the ring-grazing northern upward current regions (37).

Additional information on the correspondence between auroral emissions and upward currents is apparent in Fig. 5, which maps the spatial distribution of SKR sources during event S_3, along with the northern UV aurorae simultaneously observed by HST (29) on 2017-066 (17 March) over 15:27-16:21 UT (corrected for the light travel time). Figure 5A displays the distribution of SKR emissions observed over the same time interval as in Fig. 1, as seen from Cassini. As expected for cyclotron emission, SKR

lower- (higher-) frequency sources are observed at larger (closer) distances from the planet. Figure 5B displays in a polar view the magnetic footprint of radio sources visible in Fig. 5A. The R-X mode footprints map a circumpolar auroral oval visible from 04:00 to 22:00 LT. Its latitudinal width of about 8° to 10° is dominated by the scattering of goniopolarimetric results. This longitudinally extended range derives from the LT range swept by the spacecraft during the investigated interval. Movie 1 displays an animation

Fig. 5. Spatial distribution of SKR sources and UV aurorae. (**A**) 2D locus of SKR sources in the sky plane as seen from Cassini, from RPWS-HFR three-antenna measurements of **k** directions (*2*) on 2017-066 13:00–17:00 UT (event S_3). The applied data selection was identical to that used in Fig. 4A, except that $SNR_{\pm X,Z} > 40$ dB for clarity. The color scale and definition of symbols are identical to those of Fig. 4A. The red planetary meridian indicates noon. The typical 2° uncertainty on **k** measurements is displayed by the bottom-right ellipse. (**B**) Polar view of the magnetic footprint of radio sources displayed in (A) as a function of planetocentric latitude and LT. Error bars are not represented for clarity but are similar to those studied in (*12, 32*). The dotted-dashed curved lines display radio horizons at extremal frequencies, beyond which a radio source cannot be observed. The double green arrow indicates the Cassini LT range swept during the interval. The gray line plots the associated Cassini magnetic footprint, whose orange and green portions indicate the crossing of SKR source and of the upward current region identified in Fig. 1. (**C**) Same as (B) but for the SKR source observed from 09:00 to 09:20 UT between 20 and 27 kHz. The data selection was slightly relaxed with $T > 0.85$, $|z_r| < 0.2$ and $SNR_{\pm X,Z} > 20$ dB to maximize statistics. Cassini's LT is indicated by a green dotted-dashed arrow. The radio source footprints coincide with the crossed SKR source (orange portion of the gray line). (**D**) Polar view of Saturn's H_2 aurorae observed with the Hubble Space Telescope on day 2017-066 from 15:27 to 16:21 UT, after

correcting for light travel time. The background-subtracted image was projected onto the 1100-km-altitude surface as a function of planetocentric coordinates and expressed in kilo-rayleighs (kR) of H_2 (*2*). The gray line and its colored portions are identical to those of (B) and (C). The upward current layer (green) matches the main auroral oval, while the SKR source crossed (orange) only matches the poleward portion of it.
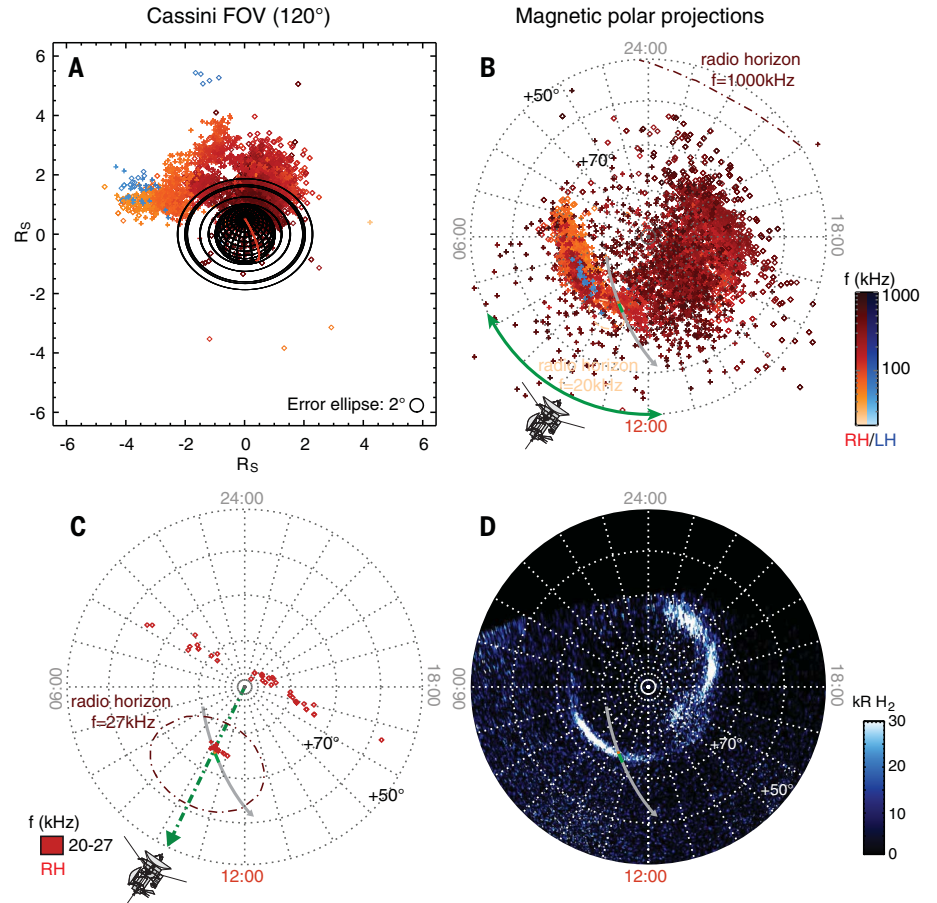
of Fig. 5, A and B, and illustrates how the spatial distribution of detected sources varies with the time-variable position of the spacecraft. The L-O mode footprints are well clustered and precisely colocated with the R-X mode footprints between 08:00 and 10:00 LT. The comparison with Fig. 5D illustrates that SKR sources are colocated with the UV auroral oval (see fig. S5 for a superimposition of both panels). This correspondence is visible both on the dawn-side sector, where the UV oval is thin and well defined between 06:00 and 12:00 LT from +72° to +75° latitudes, and on the dusk-side sector, where the UV aurorae are more spread in latitude between 14:00 and 23:00 LT and include a poleward arc beyond +80° between 14:00 and 18:00 LT. The Cassini footprint trajectory is displayed by the gray arrow in Fig. 5, B to D, over the 13:00–17:00 UT interval. Its colored portions indicate the upward current region and the SKR source region identified in Fig. 1. The upward current layer matches the main UV auroral oval in size, while the SKR source was traversed along a smaller, poleward, portion of it. This can be better seen in Fig. 5C, which is similar to Fig. 5B except that it only plots local radio sources,
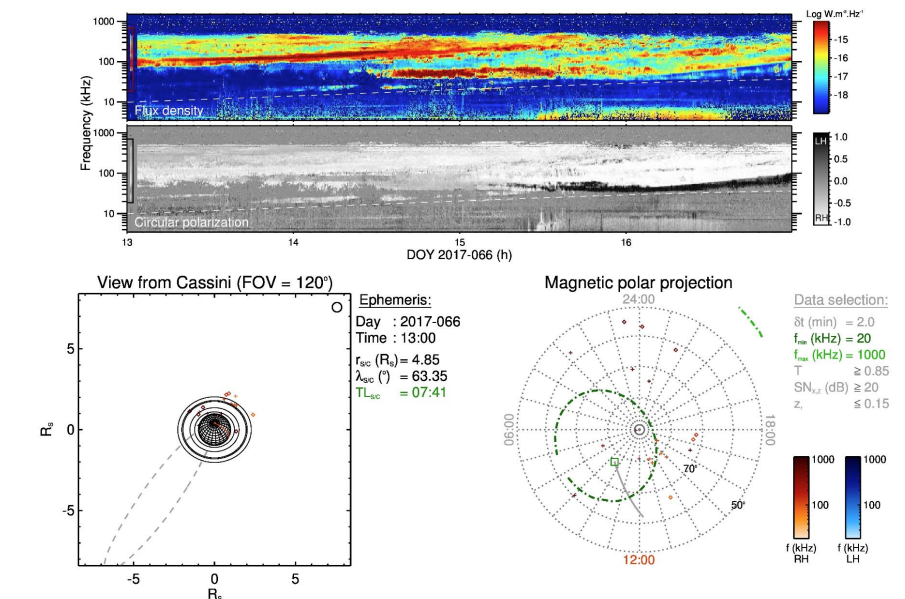
Movie 1. Animated view of the spatial distribution of SKR sources in a format identical to that of Fig. 5, A and B. The only differences are the exposure time, here fixed to 2 min for each frame, and the data selection criteria, here slightly relaxed with respect to Fig. 5, A and B, to maximize statistics (*2*).
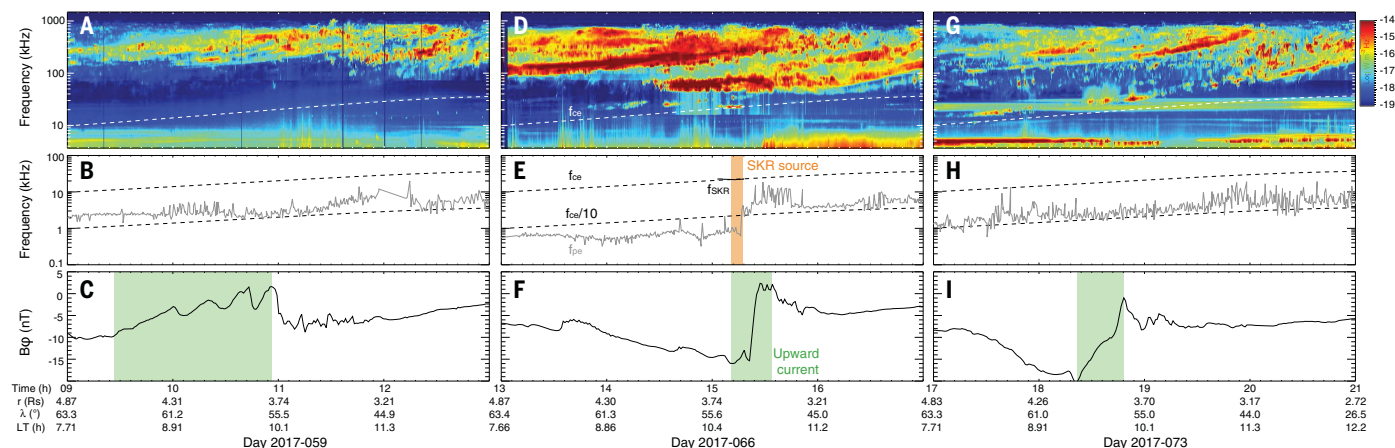
Fig. 6. Cassini radio, plasma, and magnetic measurements acquired during three successive auroral passes with similar trajectories. (**D** to **F**) Identical to Fig. 1, A, D, and E, covering the interval 2017-066 from 13:00 to 17:00 UT encompassing a SKR source crossing (orange-shaded region). (**A** to **C**) and (**G** to **I**) Same as (D) to (F) but for the intervals 2017-059 09:00-13:00 UT and 2017-073 17:00-21:00 UT (i.e., the orbit before and after), respectively. No SKR source was detected during these intervals, for which f_{pe} was always larger than $f_{ce}/10$.

radiated near f_{ce} (2). Again, the footprints of R-X mode located sources map to the Cassini trajectory and only to the inward portion of the main oval and associated upward current layer.

Implications for the generation of Saturn's auroral radio emissions

We now discuss the average properties of SKR LF emissions as a whole, in the CMI frame. The identification of only three SKR sources in the dawn-side sector and none in the dusk-side sector brings a direct in situ confirmation to the LT dependence of SKR spectrum (here studied during quiet solar wind conditions), with lowest frequencies at dawn (16, 24). The small number of radio sources identified along orbits with very similar trajectories and their limited extent additionally indicate time-variable generation conditions.

In all three cases, the sampled R-X mode emission was found to be amplified nearly perpendicularly to the local magnetic field, with emission frequency clearly below f_{ce} supplied by CMI-unstable electrons of 6- to 12-keV mean energies. Direct perpendicular emission through CMI requires specific shell-type distribution functions of the accelerated electrons in the velocity plane, such as those already measured for the two southern SKR sources crossed in 2008 (9, 13). The situation at Saturn thus directly compares to the terrestrial case, where AKR sources are commonly found to be associated with trapped, shell- or horseshoe, or ring distributions (3, 38, 39). It differs from the situation at Jupiter, which just started to be investigated in situ by the Juno spacecraft, where the hectometer and decameter radio emissions instead appear to be produced at oblique angles from the local magnetic field by loss cone electron distribution functions (40, 41). The presence of 6- to 12-keV electron beams with such distribution functions at 3.5 to 4.5 R_S distances from Saturn's center, or equivalently 2.8 to 3.7 R_S above the atmosphere along flux tubes (Table 1), implies that

at least part of the acceleration by upward currents occurred farther along the field lines, which constrains models of current-voltage relationships (42).

The general agreement with the CMI mechanism relies on fulfilling both CMI conditions. The three SKR sources encountered by Cassini matched upward currents with current densities larger than the average (37), consistent with an acceleration region extending to higher altitudes and thus satisfying CMI condition (i). Then, while the spacecraft entered the northern SKR source region when it encountered an upward current region, it exited it when f_{pe} suddenly rose, exceeding the typical $f_{ce}/10$ threshold. After exiting the source, the spacecraft remained for some time inside the upward current region, which matches the main UV oval across the Cassini trajectory. This suggests that at 3.5 to 4.5 R_S distances, SKR emission is primarily controlled by the electron plasma density N through CMI condition (ii), which quenches the radiation when f_{pe}/f_{ce} typically exceeds 0.1. According to the Knight relationship (43) and assuming a constant current density, larger N is also expected to decrease electron acceleration and weaken CMI condition (i).

Figure 6 shows radio, plasma, and magnetic measurements for three successive, quasi-identical, passes of the spacecraft into the northern dawn-side auroral region. Figure 6, D to F, correspond to event S₃ (orbit 264), while Fig. 6, A to C and G to I, correspond to the previous (orbit 263) and the following (orbit 265) auroral passes. The spacecraft did not fly across any local SKR source during orbits 263 and 265, while it encountered regions of upward current in both cases with current densities of 29 and 108 nA m⁻². On orbit 265, the SKR intensity was relatively high and its spectrum reached frequencies as low as 20 kHz, while the upward current density was as strong as that of orbit 264. Instead, on orbit 263, SKR was comparatively faint, with a spectrum confined to higher frequencies and a lower upward

current density. This suggests that CMI condition (i), which requires unstable electrons at 3.6 R_S, was at least satisfied for orbit 265. For both intervals, however, f_{pe} was systematically larger than $f_{ce}/10$ within the upward current region, in contrast with orbit 264 (event S₃), indicating that the mean level of N over hours-long intervals strongly varies with time from pass to pass. These results thus indicate that CMI condition (ii) was not fulfilled for orbits 263 and 265 and likely quenched SKR LF emission during orbit 265.

The origin of time-variable magnetospheric plasma densities at the altitude of LF SKR sources has now to be elucidated. The asymmetrical plasmapause-like boundary, a transition from high to low plasma densities with increasing latitudes that rotates at the northern SKR period, is a plausible source for time-variable N (44). On a shorter time scale, Fig. 1 does not display any Earth-like plasma cavity, as already noticed for the 2008-291 southern event (4). However, Fig. 2 (and possibly fig. S2) shows small-sized, recurrent, drops of f_{pe} colocated with SKR sources for event S₁ (and S₂). This suggests that acceleration features could coincide with localized radio sources, as at Earth. Definite conclusions, however, rely on knowing whether the plasma region was dominated by hot electrons (such as on day 2008-291) or could have included a cold plasma component sensitive to local potential drops and able to map plasma cavities. As the CAPS experiment was not operating during these orbits, these issues remain unresolved.

Finally, the very narrow-banded nature of SKR sources, particularly obvious during event S₁ (Fig. 3B), is consistent with either a nonsaturated linear wave growth, as predicted by the CMI theory and measured in laboratory [e.g., (45)], or with saturation by coherent nonlinear trapping, previously assumed in models of the SKR spectrum (46). However, the emission levels may be too low to reach saturation, and the observed wave intensities (Table 1) are also consistent with linear growth.

REFERENCES AND NOTES

1. L. Lamy, The Saturnian Kilometric Radiation before the Cassini Grand Finale, in *Planetary Radio Emissions VIII*, G. Fischer, G. Mann, M. Panchenko, P. Zarka, Eds. (Austrian Academy of Sciences Press, Vienna, 2017), pp. 171–190.
2. Materials and methods are available as supplementary materials.
3. R. A. Treumann, The electron-cyclotron maser for astrophysical applications. *Astron. Astrophys. Rev.* **13**, 229–315 (2006). doi: 10.1007/s00159-006-0001-y
4. L. Lamy *et al.*, Properties of Saturn kilometric radiation measured within its source region. *Geophys. Res. Lett.* **37**, L12104 (2010). doi: 10.1029/2010GL043415
5. E. J. Bunce *et al.*, Extraordinary field-aligned current signatures in Saturn's high-latitude magnetosphere: Analysis of Cassini data during revolution 89. *J. Geophys. Res.* **115**, A10238 (2010). doi: 10.1029/2010JA015612
6. W. S. Kurth *et al.*, A close encounter with a Saturn kilometric radiation source region, in *Planetary Radio Emissions VII*, H. O. Rucker, W. S. Kurth, P. Louarn, G. Fischer, Eds. (Austrian Academy of Sciences Press, Vienna, 2011), pp. 75–85.
7. P. Schippers *et al.*, Auroral electron distributions within and close to the Saturn kilometric radiation source region. *J. Geophys. Res.* **166**, A05203 (2011). doi: 10.1029/2011JA016461
8. L. Lamy *et al.*, Emission and propagation of Saturn kilometric radiation: Magnetoionic modes, beaming pattern, and polarization state. *J. Geophys. Res.* **116**, A04212 (2011). doi: 10.1029/2010JA016195
9. R. L. Mutel *et al.*, CMI growth rates for Saturnian kilometric radiation. *Geophys. Res. Lett.* **37**, L19105 (2010). doi: 10.1029/2010GL044940
10. A. Hilgers, The auroral radiating plasma cavities. *Geophys. Res. Lett.* **19**, 237–240 (1992). doi: 10.1029/91GL02938
11. J. D. Menietti *et al.*, Saturn kilometric radiation near a source center on day 73. In *Planetary Radio Emissions VII*, H.O. Rucker, W. S. Kurth, P. Louarn, Eds. (Austrian Academy of Sciences Press, Vienna, 2011), pp. 87–95.
12. L. Lamy *et al.*, An auroral oval at the footprint of Saturn's kilometric radio sources, colocated with the UV aurorae. *J. Geophys. Res.* **114**, A10212 (2009). doi: 10.1029/2009JA014401
13. J. D. Menietti *et al.*, Analysis of Saturn kilometric radiation near a source center. *J. Geophys. Res.* **116**, A12222 (2011). doi: 10.1029/2011JA017056
14. W. S. Kurth *et al.*, An Earth-like correspondence between Saturn's auroral features and radio emission. *Nature* **433**, 722–725 (2005). doi: 10.1038/nature03334; pmid: 15716947
15. W. S. Kurth *et al.*, Saturn kilometric radiation intensities during the Saturn auroral campaign of 2013. *Icarus* **263**, 2–9 (2016). doi: 10.1016/j.icarus.2015.01.003
16. L. Lamy *et al.*, Saturn kilometric radiation: Average and statistical properties. *J. Geophys. Res.* **113** (A7), A07201 (2008). doi: 10.1029/2007JA012900
17. S. V. Badman, S. W. H. Cowley, L. Lamy, B. Cecconi, P. Zarka, Relationship between solar wind corotating interaction regions and the phasing and intensity of Saturn kilometric radiation bursts. *Ann. Geophys.* **26**, 3641–3651 (2008). doi: 10.5194/angeo-26-3641-2008
18. D. A. Gurnett *et al.*, Discovery of a north-south asymmetry in Saturn's radio rotation period. *Geophys. Res. Lett.* **36**, L16102 (2009). doi: 10.1029/2009GL039621
19. D. G. Mitchell *et al.*, Recurrent energization of plasma in the midnight-to-dawn quadrant of Saturn's magnetosphere, and its relationship to auroral UV and radio emissions.

Planet. Space Sci. **57**, 1732–1742 (2009). doi: 10.1016/j.pss.2009.04.002
20. C. M. Jackman *et al.*, On the character and distribution of lower-frequency radio emissions at Saturn and their relationship to substorm-like events. *J. Geophys. Res.* **114**, A08211 (2009). doi: 10.1029/2008JA013997
21. C. M. Jackman *et al.*, In situ observations of the effect of a solar wind compression on Saturn's magnetotail. *J. Geophys. Res.* **115**, A10240 (2010). doi: 10.1029/2010JA015312
22. L. Lamy, Variability of southern and northern SKR periodicities, in *Planetary Radio Emissions VII*, H.O. Rucker, W.S. Kurth, P. Louarn, G. Fischer, Eds. (Austrian Academy of Sciences Press, Vienna, 2011), pp. 39–50.
23. L. Lamy *et al.*, Multispectral simultaneous diagnosis of Saturn's aurorae throughout a planetary rotation. *J. Geophys. Res.* **118**, 4817–4843 (2013). doi: 10.1002/jgra.50404
24. T. Kimura *et al.*, Long-term modulations of Saturns auroral radio emissions by the solar wind and seasonal variations controlled by the solar ultraviolet flux. *J. Geophys. Res.* **118**, 7019–7035 (2013). doi: 10.1002/2013JA018833
25. J. J. Reed, C. M. Jackman, L. Lamy, W. S. Kurth, D. K. Whiter, Low Frequency Extensions of the Saturn Kilometric Radiation as a proxy for magnetospheric dynamics. *J. Geophys. Res.* **123**, 443–463 (2018). doi: 10.1002/2017JA024499
26. D. A. Gurnett *et al.*, The Cassini Radio and Plasma Wave Investigation. *Space Sci. Rev.* **114**, 395–463 (2004). doi: 10.1007/s11214-004-1434-0
27. M. K. Dougherty *et al.*, The Cassini magnetic field investigation. *Space Sci. Rev.* **114**, 331–383 (2004). doi: 10.1007/s11214-004-1432-2
28. D. T. Young *et al.*, Cassini Plasma Spectrometer investigation. *Space Sci. Rev.* **114**, 1–112 (2004). doi: 10.1007/s11214-004-1406-4
29. L. Lamy *et al.*, Saturn's northern aurorae at solstice from HST observations coordinated with Cassini's Grand Finale. *Geophys. Res. Lett.* **45**, (2018). doi: 10.1029/2018GL078211
30. W. S. Kurth *et al.*, High spectral and temporal resolution observations of Saturn kilometric radiation. *Geophys. Res. Lett.* **32**, L20S07 (2005). doi: 10.1029/2005GL022648
31. J. D. Kraus, Wave polarization, in *Radio Astronomy*, 116–125 (McGraw-Hill, New York, 1966).
32. B. Cecconi *et al.*, Goniopolarimetric study of the revolution 29 perikrone using the Cassini Radio and Plasma Wave Science instrument high-frequency radio receiver. *J. Geophys. Res.* **114**, A03215 (2009). doi: 10.1029/2008JA013830
33. L. Lamy, P. Zarka, B. Cecconi, S. Hess, R. Prangé, Modeling of Saturn kilometric radiation arcs and equatorial shadow zone. *J. Geophys. Res.* **113** (A10), A10213 (2008). doi: 10.1029/2008JA013464
34. R. L. Mutel, I. W. Christopher, J. S. Pickett, Cluster multispacecraft determination of AKR angular beaming. *Geophys. Res. Lett.* **37**, L07104 (2008). doi: 10.1029/2008GL033377
35. P. H. M. Galopeau, M. Y. Boudjada, An oblate beaming cone for Io-controlled Jovian decameter emission. *J. Geophys. Res.* **121**, 3120–3138 (2016). doi: 10.1002/2015JA021038
36. D. B. Melrose, G. A. Dulk, On the elliptical polarization of Jupiter's decametric radio emission. *Astron. Astrophys.* **249**, 250–257 (1991).
37. G. J. Hunt *et al.*, Field-aligned currents in Saturn's magnetosphere: Observations from the F-ring orbits. *J. Geophys. Res.* **45**, 4692–4700 (2018). doi: 10.1029/2018GL077925
38. P. Louarn *et al.*, Trapped electrons as a free energy source for the auroral kilometric radiation. *J. Geophys. Res.* **95** (A5), 5983–5995 (1990). doi: 10.1029/JA095iA05p05983

39. R. E. Ergun *et al.*, Electron-Cyclotron Maser Driven by Charged-Particle Acceleration from Magnetic Fieldaligned Electric Fields. *Astrophys. J.* **538**, 456–466 (2000). doi: 10.1086/309094
40. P. Louarn *et al.*, Generation of the Jovian hectometric radiation: First lessons from Juno. *Geophys. Res. Lett.* **44**, 4439–4446 (2017). doi: 10.1002/2017GL072923
41. C. K. Louis *et al.*, Io-Jupiter decametric arcs observed by Juno/Waves compared to ExPRES simulations. *Geophys. Res. Lett.* **44**, 9225–9232 (2017). doi: 10.1002/2017GL073036
42. L. C. Ray, M. Galand, P. A. Delamere, B. L. Fleshman, Current-voltage relation for the Saturnian system. *J. Geophys. Res. Space Phys.* **118**, 3214–3222 (2013). doi: 10.1002/jgra.50330
43. S. Knight, Parallel electric fields. *Planet. Space Sci.* **21**, 741–750 (1973). doi: 10.1016/0032-0633(73)90093-7
44. D. A. Gurnett *et al.*, The rotation of the plasmapause-like boundary at high latitudes in Saturn's magnetosphere and its relation to the eccentric rotation of the northern and southern auroral ovals. *Geophys. Res. Lett.* **38**, L21203 (2011). doi: 10.1029/2011GL049547
45. R. Bingham *et al.*, Laboratory astrophysics: Investigation of planetary and astrophysical maser emission. *Space Sci. Rev.* **178**, 695–713 (2013). doi: 10.1007/s11214-013-9963-z
46. P. H. M. Galopeau, P. Zarka, D. Le Quéau, Theoretical model of Saturn's kilometric radiation spectrum. *J. Geophys. Res.* **94** (A7), 8739–8755 (1989). doi: 10.1029/JA094iA07p08739

RESEARCH ARTICLE SUMMARY

GAS GIANT PLANETS

In situ collection of dust grains falling from Saturn's rings into its atmosphere

Hsiang-Wen Hsu*, Jürgen Schmidt, Sascha Kempf, Frank Postberg, Georg Moragas-Klostermeyer, Martin Seiß, Holger Hoffmann, Marcia Burton, ShengYi Ye, William S. Kurth, Mihály Horányi, Nozair Khawaja, Frank Spahn, Daniel Schirdewahn, James O'Donoghue, Luke Moore, Jeff Cuzzi, Geraint H. Jones, Ralf Srama

INTRODUCTION: During the Cassini spacecraft's Grand Finale mission in 2017, it performed 22 traversals of the 2000-km-wide region between Saturn and its innermost D ring. During these traversals, the onboard cosmic dust analyzer (CDA) sought to collect material released from the main rings. The science goals were to measure the composition of ring material and determine whether it is falling into the planet's atmosphere.

RATIONALE: Clues about the origin of Saturn's massive main rings may lie in their composition. Remote observations have shown that they are formed primarily of water ice, with small amounts of other materials such as silicates, complex organics, and nanophase hematite. Fine-grain ejecta generated by hypervelocity collisions of interplanetary dust particles (IDPs) on the main rings serve as microscopic samples. These grains could be examined in situ by the Cassini spacecraft during its final orbits. Deposition of ring ejecta into Saturn's atmosphere has been suggested as an explanation for the pattern of ionospheric H_3^+ infrared emission, a phenomenon known as ring rain. Dynamical studies have suggested a preferential transport of charged ring particles toward the planet's southern hemisphere because of the northward offset of Saturn's internal magnetic field. However, the deposition flux and its form (ions or charged grains) remained unclear. In situ characterization of the ring ejecta by the Cassini CDA was planned to provide observational constraints on the composition of Saturn's ring system and test the ring rain hypothesis.

RESULTS: The region within Saturn's D ring is populated predominantly by grains tens of nanometers in radius. Larger grains (hundreds of nanometers) dominate the mass density but are narrowly confined within a few hundred kilometers around the ring plane. The measured flux profiles vary with the CDA pointing configurations. The highest dust flux was registered during the ring plane crossings when the CDA was sensitive to the prograde dust populations (Kepler ram pointing) (see the figure). When the CDA was pointed toward the retrograde direction (plasma ram pointing), two additional flux enhancements appeared on both sides of the rings at roughly the same magnetic latitude. The south dust peak is stronger and wider, indicating the dominance of Saturn's magnetic field in the dynamics of charged nanograins. These grains are likely fast ejecta released from the main rings and falling into Saturn, producing the observed ionospheric signature of ring rain.

We estimate that a few tons of nanometer-sized ejecta is produced each second across the main rings. Although this constitutes only a small fraction (<0.1%) of the total ring ejecta production, it is sufficient to supply the observed ring rain effect. Two distinct grain compositional types were identified: water ice and silicate. The silicate-to-ice ratio varies with latitude; the global average ranges from 1:11 to 1:2, higher than that inferred from remote observations of the rings.

CONCLUSION: Our observations illustrate the interactions between Saturn and its main rings through charged, nanometer-sized ejecta particles. The dominance of nanograins between Saturn and its rings is a dynamical selection effect, stemming from the grains' high ejection speeds (hundreds of meters per second and higher) and Saturn's offset magnetic field. The presence of the main rings modifies the effects of the IDP infall to Saturn's atmosphere. The rings do this asymmetrically, leading to the distribution of the ring rain phenomenon. Confirmed ring constituents include water ice and silicates, whose ratio is likely shaped by processes associated with ring erosion processes and ring-planet interactions. ∎
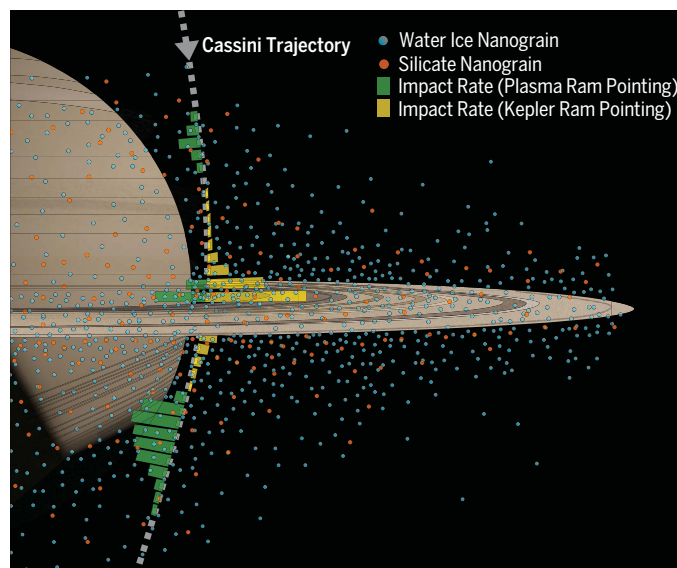
Schematic view of the nanometer-sized ring ejecta environment in the vicinity of Saturn. CDA measurements were taken during Cassini's Grand Finale mission. The measured dust flux profiles, presented by the histograms along the spacecraft trajectory, show different patterns depending on the instrument pointing configuration. The highest dust flux occurred at the ring plane under Kepler ram pointing (yellow). The profiles registered with plasma ram pointing (green) show two additional, mid-latitude peaks at both sides of the rings with substantial north-south asymmetry. This signature in the vertical profiles indicates that the measured nanograins in fact originate from the rings and are whirling into Saturn under the dynamical influence of the planet's offset magnetic field. Blue and orange dots represent the two grain composition types identified in the mass spectra, water ice and silicate, respectively.

RESEARCH ARTICLE

GAS GIANT PLANETS

In situ collection of dust grains falling from Saturn's rings into its atmosphere

Hsiang-Wen Hsu[1]*, Jürgen Schmidt[2], Sascha Kempf[1], Frank Postberg[3,4],
Georg Moragas-Klostermeyer[5], Martin Seiß[6], Holger Hoffmann[6], Marcia Burton[7],
ShengYi Ye[8], William S. Kurth[8], Mihály Horányi[1], Nozair Khawaja[3,4], Frank Spahn[6],
Daniel Schirdewahn[6], James O'Donoghue[9], Luke Moore[10], Jeff Cuzzi[11],
Geraint H. Jones[12,13], Ralf Srama[5,14]

Saturn's main rings are composed of >95% water ice, and the nature of the remaining few percent has remained unclear. The Cassini spacecraft's traversals between Saturn and its innermost D ring allowed its cosmic dust analyzer (CDA) to collect material released from the main rings and to characterize the ring material infall into Saturn. We report the direct in situ detection of material from Saturn's dense rings by the CDA impact mass spectrometer. Most detected grains are a few tens of nanometers in size and dynamically associated with the previously inferred "ring rain." Silicate and water-ice grains were identified, in proportions that vary with latitude. Silicate grains constitute up to 30% of infalling grains, a higher percentage than the bulk silicate content of the rings.

Infrared and radio observations (*1*–*5*) have shown that Saturn's A and B rings are composed mostly (95 to 99%) of water ice, whereas in the less dense C ring and the Cassini Division the non-icy component can be higher, up to 10% in the C ring (*5*–*7*). Because of the rings' large surface-to-mass ratio, the evolution of the ring composition is dominated by the infall of material from interplanetary space, which is composed primarily of silicates and carbon-rich organics (*8*). The radial density profile of the rings is shaped by various dynamical processes, including viscous spreading, resonances with satellites, intrinsic instabilities, and ballistic transport (*5*, *9*–*13*) of ejecta released by impacts of interplanetary dust particles on the rings (*14*). The latter process is primarily responsible for redistributing and mixing the icy and non-icy material across the rings, as well as for the formation of sharp inner A and B ring edges (*10*, *15*).

A fraction of the material released by erosion from the rings is expected to fall into Saturn's atmosphere as "ring rain" (*16*–*18*). Charged particles from the rings, in the form of ions or nanoparticles, gyrate and bounce along the planet's magnetic field lines and preferentially migrate into Saturn's southern atmosphere (*19*–*25*). However, whether the infalling mass flux is sufficient to drive the observed atmospheric processes and whether the infall occurs in the form of ions or dust remain unclear.

We report the in situ detection of nanograins originating from Saturn's main rings by the cosmic dust analyzer (CDA) (*26*) during Cassini's traversals through the region between the atmosphere and the innermost D ring [1.11 to 1.24 Saturn radii (R_S) = 60,268 km], as shown in Fig. 1. Previously, Cassini instruments observed nanograins that were emerging from the E ring region and evolving into interplanetary space (*27*–*29*) and that were directly within the Enceladus ice particle plume (*30*). From the 22 spacecraft crossings of the inner D ring region during Cassini's Grand Finale mission phase (26 April 2017 to 15 September 2017), eight orbits were useful for CDA measurements because of the optimized instrument pointing (table S1). The spacecraft speed relative to the dust grains was more than 30 km s^{-1} around the closest approach (periapsis). The CDA mass analyzer (MA), a linear impact time-of-flight mass spectrometer, provides elemental compositional information for grains with radii in excess of a few nanometers striking the detector (*26*). The CDA's high rate detector (HRD), a foil detector for characterizing dense dust environments (*26*), measured the flux of grains in excess of 600 nm in radius during the traversals.

Grain composition

The MA recorded more than 2700 impact mass spectra during the eight inner D ring traversals favorable for CDA measurements. Of these spectra, 78% are very faint and at most show only peaks in the mass spectra associated with the chemical analyzer target (CAT) (made of rhodium) and its known contaminants, Rh$^+$, C$^+$, Na$^+$, and K$^+$ (*31*). These spectra do not provide conclusive compositional information about the impacting particle but do provide evidence for the number of impacts by the smallest detectable grains. The remaining 22% of the spectra exhibit peaks associated with particle constituents with sufficient signal-to-noise ratio for a compositional analysis. We identified two distinct compositional types: water ice and silicate (Fig. 2). Ice grains are identified by the presence of at least two of the following water cations: O$^+$, OH$^+$, H$_2$O$^+$, and H$_3$O$^+$. Identification of silicate grains requires a mass peak at 28 u (where u is the atomic mass unit), consistent with Si$^+$, and at least one additional mass line (Mg$^+$, Ca$^+$, or Fe$^+$) from metallic cations other than Na$^+$ and K$^+$. In total, we identified 422 water-ice–type and 214 silicate-type particles, all recorded around the passage of the ring plane within ±50° latitude (see Fig. 1 for the geometry). We did not identify any nanograins composed predominantly of iron, iron oxide, or an organic material. Because of the expected fractionation of larger organic molecules at the high impact velocities during these measurements, low-mass organic fragment cations at 12 u (C$^+$) or 13 u (CH$^+$) might overlap with the carbon contamination of the CDA impact target (*31*), curtailing our capability to identify minor organic constituents. The impact charge signals of the water-ice and silicate particles are similar in both amplitude and distribution (see supplementary text). On the basis of the impact charge, which depends linearly on the grain mass for collisions at the same speed (*32*), the majority of the detected grains are smaller than 50 nm. Taking into account the material-dependent impact charge yield, the CDA target contamination, and simulations of the dynamics of these small grains, we conclude that the ice-to-silicate number ratio of 2:1 translates into a mass ratio of the ice and silicate grains between 2:1 and 11:1 (see supplementary text). The overall silicate nanograin mass fraction as seen by the CDA interior to the D ring ranges from 8 to 30%, which is higher than the concentration of non-ice components in the rings inferred from either optical or microwave measurements (*6*, *7*, *33*).

Spatial distribution

The ring plane crossings of the 22 Grand Finale orbits occurred near local noon. The spacecraft always approached the rings from the north and reached periapsis about 10 min after crossing the ring plane at a latitude of 6° south (Fig. 1). The

[1]Laboratory for Atmospheric and Space Physics, University of Colorado–Boulder, Boulder, CO, USA. [2]Astronomy Research Unit, University of Oulu, Oulu, Finland. [3]Institut für Geowissenschaften, Universität Heidelberg, Heidelberg, Germany. [4]Institut für Geologische Wissenschaften, Freie Universität Berlin, Berlin, Germany. [5]Institut für Raumfahrtsysteme, Universität Stuttgart, Stuttgart, Germany. [6]Institut für Physik und Astronomie, Universität Potsdam, Potsdam, Germany. [7]Jet Propulsion Laboratory, Pasadena, CA, USA. [8]Department of Physics and Astronomy, University of Iowa, Iowa City, IA, USA. [9]NASA Goddard Space Flight Center, Greenbelt, MD, USA. [10]Center for Space Physics, Boston University, Boston, MA, USA. [11]NASA Ames Research Center, Moffett Field, CA, USA. [12]Mullard Space Science Laboratory, University College London, Holmbury St. Mary, Dorking, UK. [13]The Centre for Planetary Sciences at University College London/Birkbeck, London, UK. [14]Center for Astrophysics, Space Physics, and Engineering Research, Baylor University, Waco, TX, USA.
*Corresponding author. Email: sean.hsu@lasp.colorado.edu

orbits also traversed the magnetic field lines connecting the planet to its main rings when the spacecraft was within ±50° from the ring plane. The high dust impact speed of ~30 km s^{-1} enabled the MA to detect grains as small as tens of nanometers. In contrast, the HRD was sensitive to grains with radii larger than 600 nm. Overall, the CDA was sensitive to dust impacts when the instrument was oriented either toward the Kepler ram direction (i.e., the impact direction of grains moving in circular bound, prograde orbits) or toward the plasma ram direction (i.e., the flow direction of plasma co-rotating with Saturn's magnetic field). In total, the CDA obtained dust density profiles during five orbits with Kepler ram and three orbits with plasma ram pointing (see supplementary text). Because the CDA was not calibrated for such high impact speeds, the grain size and impact speed cannot be directly derived from the recorded impact charge waveforms (27). We corrected the MA detection rate, which was frequently close to saturation (approximately one impact per second), for instrumental dead-time effects (34).

The MA detections reveal that the 2000-km-wide gap between Saturn's cloud tops and the inner rim of the D ring is populated by grains a few tens of nanometers in radii. The MA impact rate profiles strongly depend on the instrument orientation (Fig. 3, A and B). We observed the impact rate maximum at the time of the ring plane crossing during all CDA orbits with Kepler ram instrument orientation. The observed peak rates between 7 and 17 impacts per second correspond to a nanograin number density of 3.5 × 10^{-2} to 8.5 × 10^{-2} m^{-3}.

The HRD foil detector was sensitive to larger grains moving in Keplerian orbits and registered 13 particles of >600 nm close to the ring plane, which corresponds to a number density of (2 ± 1) × 10^{-4} m^{-3} (Fig. 3C). The Cassini Radio and Plasma Wave Science instrument (RPWS) (35), which was sensitive to plasma signals created by dust impacts onto the spacecraft, had a similar-size-threshold HRD (36). The full widths at half maximum (FWHMs) of the diffuse dust ring interior to the D ring derived from RPWS (37) and HRD data are 300 and 900 km, respectively (Fig. 3C). RPWS provided a more detailed ring structure profile because of its higher temporal resolution, which explains the differences in the derived ring thickness. Given the instrument dead-time effects (34), MA is biased against sampling the larger grains detected by the HRD and RPWS because of their lower number density. We focus on the MA nanograin measurements and their implications below, but the mass density of larger grains is one to two orders of magnitude higher.

For orbits with plasma ram orientation, the MA rate shows two broader peaks in addition to the peak in the ring plane (Fig. 3B). They are centered at roughly the same magnetic latitude north and south, corresponding to $L = 1.35$ ($L = r/R_S$, which depicts the dipolar magnetic field line piercing the planet's symmetry plane at a radial distance r), which suggests Lorentz force–driven dust dynam-

ics and the main rings as the dust source. We also observed a pronounced north-south asymmetry of the MA count rate, which we attributed to the northward offset of Saturn's magnetic dipole: The south dust peak is two to three times stronger and much broader than the north peak. Generally, the peak impact rate was an order of magnitude lower than that during Kepler ram orbits.

The fraction of ice nanograins also varies with latitude (Fig. 4A), decreasing from 70 to 90% near the ring plane to ~40% at ~30° latitude, beyond which the small number of detections is insufficient to determine the fraction. In the northern hemisphere, the nanograin density profile is similar to the H_3^+ emission derived from infrared images of Saturn's atmosphere (16–18) (Fig. 4B).
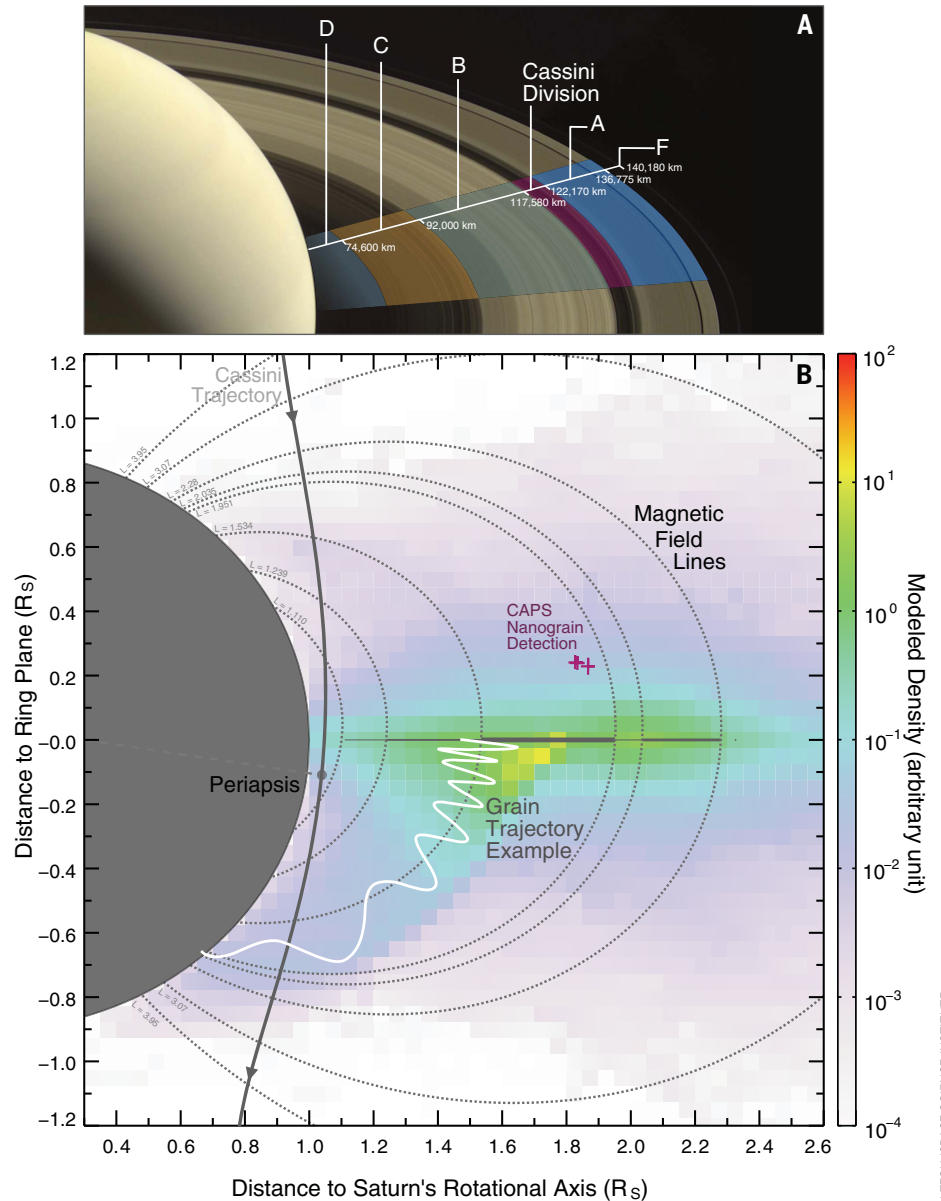
Fig. 1. Edge-on view of a typical Cassini Grand Finale orbit. (**A**) A mosaic image of Saturn's rings taken by Cassini. (**B**) The gray curve is the Cassini trajectory around its periapsis, with our model grain density profile in the background. The grid size is 0.05 R_S, corresponding to about 300 s of CDA measurements along the trajectory. Dotted curves mark planetary magnetic field lines connecting the edges of the main ring segments D, C, B, and A, as well as the orbits of icy moons Mimas ($L = 3.07$) and Enceladus ($L = 3.95$), with Saturn's atmosphere (39). Purple crosses indicate the locations of negatively charged nanograin detections by the Cassini plasma spectrometer (CAPS) (51) during Cassini's Saturn orbit insertion in 2004 near an L shell of 1.9. An example trajectory of a charged nanograin launched from the main rings (white line) demonstrates the strong influence of the Lorentz forces on the grain dynamics, which leads to the grain's deposition in Saturn's southern hemisphere.

The latitudinal H_3^+ emission pattern is indicative of ring material infalling through magnetic connection (ring rain).

Nanograin dynamics

Because the charge-to-mass ratio of the nanograins is high, their dynamics are dominated by the Lorentz force rather than gravity. Thus, a model based on gravitational accelerations alone is not adequate to describe the orbital evolution of such grains. We have developed a numerical model that follows the dynamical evolution of nanograins from their origin on the surface of the main rings, as impact ejecta produced from exogenous hypervelocity collisions, until they are lost to Saturn or the rings or they have moved beyond 2.5 R_S. The model includes relevant forces acting on charged grains: gravity, Lorentz force, and atmospheric drag (*38*). We employ the axi-symmetric Z3 model for describing Saturn's magnetic field (*39*). The instantaneous grain charge resulting from the interaction with the ambient plasma environment and solar ultraviolet radiation is allowed to vary along the grain orbit. We use a plasma model, which includes Saturn's ionosphere (*40*) and a simple ring ionosphere (*38*). By building up a large library of test particle trajectories, our model predicts the nanograin flux and number density near Saturn's main rings (Fig. 1B).

The measured number of ions created by the nanograin impacts on the CDA's rhodium target indicates that the grains have to be smaller than ~50 nm. From comparison to the dust dynamics model, we find that grains with radii of ~15 to 20 nm best reproduce the rate profiles obtained during Kepler ram and plasma ram orbits (Fig. 3) (*38*). Larger grains are less influenced by the Lorentz force so cannot reach the midlatitudes observed during the plasma ram orbits. Charged grains smaller than 15 nm, on the other hand, are dynamically dominated by the Lorentz force and would populate the entire region, which is not consistent with the observed number density enhancements at the midlatitudes. It is unlikely for the CDA to detect grains smaller than ~5 nm at an impact speed of 30 km s^{-1}. Nanograins with radii of ~20 nm move mostly in the prograde direction and are detected predominantly during Kepler ram orbits. Because of their substantial charge-to-mass ratio, the particles gyrate like an ion along the planetary magnetic field lines. Toward higher latitudes, charged grains are deflected when they reach their magnetic mirror points, where their motion parallel to the field line is reversed. Because the grains' velocity component perpendicular to the magnetic field lines increases toward their mirror points, the nanograin flux arriving from the plasma ram direction is enhanced at the midlatitudes.

The spatial distribution of the nanograins also depends strongly on the grains' initial speed—the speed at which the freshly created nanograin ejecta are jettisoned off the surfaces of the main rings. We find that the ejection speed needs to be larger than a few hundred meters per second to match the rate profiles of the Kepler ram orbits, and it needs to be several thousand meters per second to reproduce the northern peak of the plasma ram orbits. The required high ejection speeds are consistent with data from previous theoretical studies (*22*) and have been observed in laboratory experiments (impact jetting) (*41*, *42*). The ejection speed is likely to be larger for smaller ejecta (*43*, *44*), implying much higher ejection speeds for nanograins than for micrometer-sized and larger ejecta, which fuel the ballistic transport (*10*, *15*).

Nanograin origin

Our best-fitting simulations of the MA impact rate during Kepler ram orbits indicate that 40% of the detected grains originate from the optically thick B ring, 50% from the C ring, and the rest from the A and D rings, as well as the Cassini Division. The ratio between grains from the B ring and those from the C ring arriving at the Cassini's trajectory is almost constant, implying that the latitudinal variation of the ice fraction (Fig. 4A) is not due to the rings' composition. This conclusion is supported by the much higher
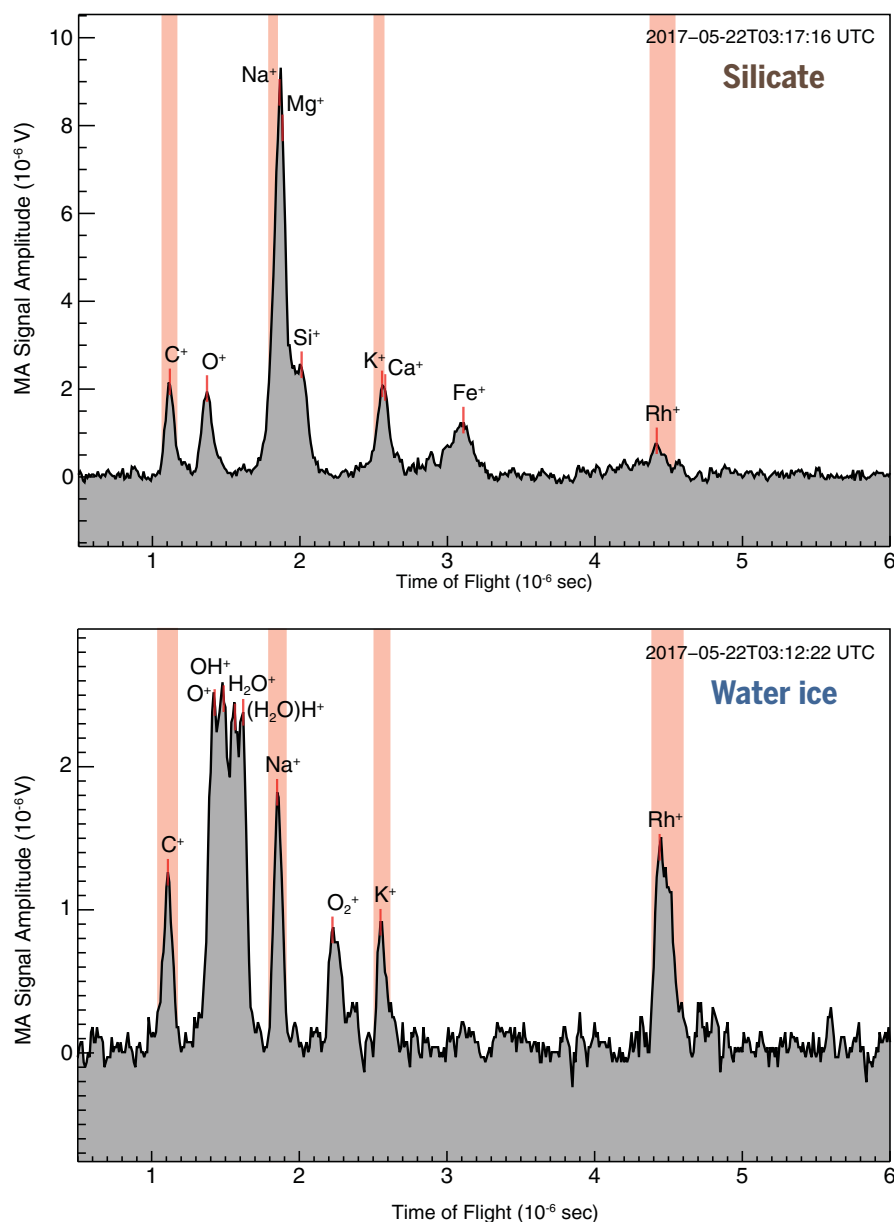
Fig. 2. Example mass spectra for the two nanograin populations detected interior to the D ring. Both mass spectra were recorded over a few minutes around the ring plane crossing of orbit 275 on 22 May 2017 (day 142). Vertical shaded regions indicate mass lines stemming from the CAT (Rh^+) and target contaminants (C^+, Na^+, and K^+). Na^+ and K^+ contamination probably arises from the remains of salt-rich grains that impacted the instrument during prior Enceladus plume traversals (*31*). However, Na^+ and K^+ cannot be completely ruled out as grain constituents. (Top) Mass spectrum of a silicate nanograin characterized by its Mg^+, Si^+, and Fe^+ mass lines. (Bottom) Mass spectrum of a water-ice nanograin characterized by its O^+, OH^+, H_2O^+, and H_3O^+ mass lines.
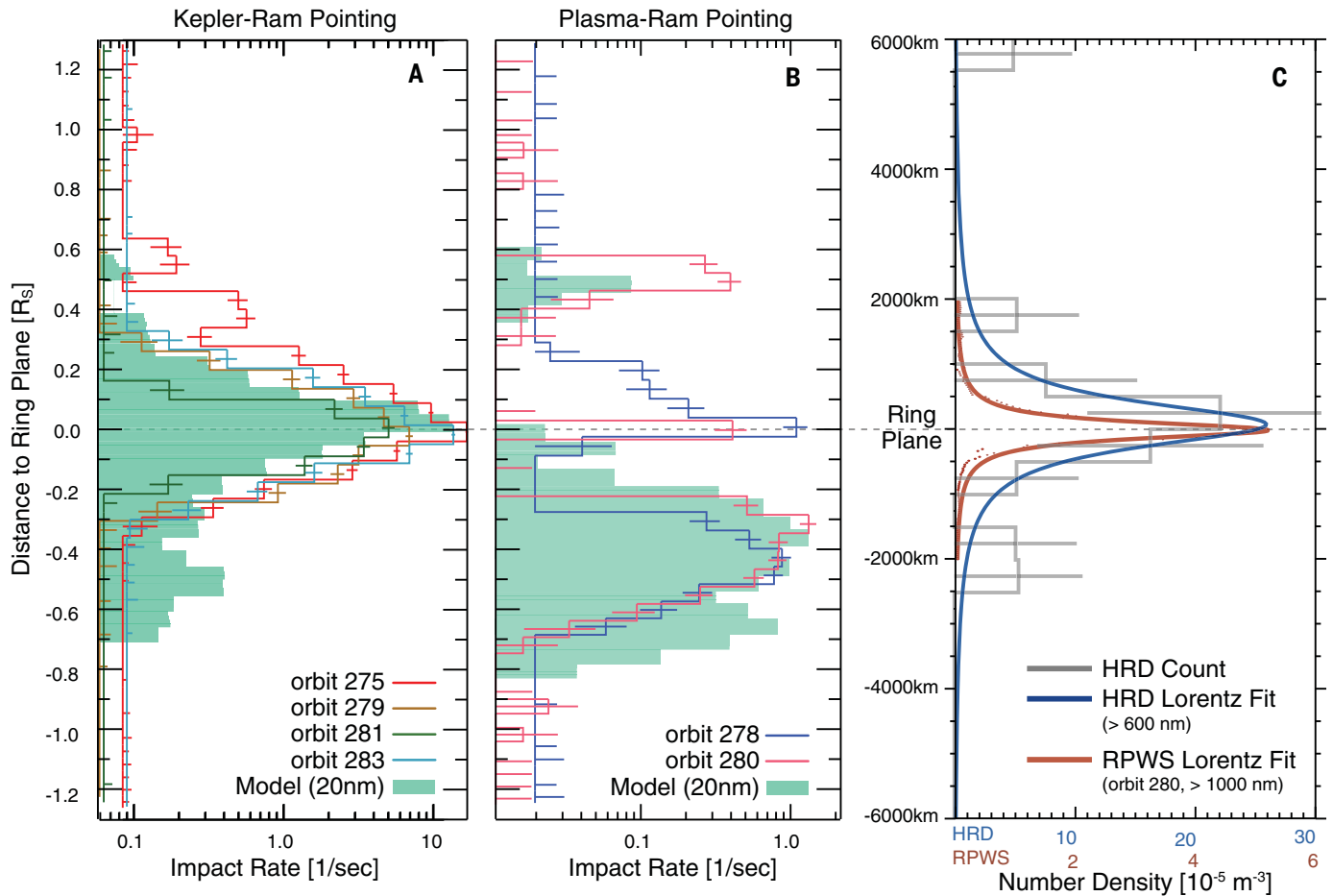
Fig. 3. Vertical density profiles measured by the CDA (MA and HRD) and RPWS. The MA impact rate is plotted as a function of the ring plane distance during (**A**) Kepler ram and (**B**) plasma ram orbits. The green histograms are the results of our model calculations for 20-nm grains. The impact rate was generally about 10 times higher during Kepler ram orbits than plasma ram orbits. (**C**) The profile of ring particles interior to the D ring, measured by the HRD and RPWS (37). The instruments had similar size detection thresholds (600 nm and 1000 nm), but RPWS had a much larger sensitive area and thus a much higher temporal resolution. Overlain are fitted Lorentz density profiles, with vertical FWHMs of 900 km (HRD) and 300 km (RPWS). The RPWS peak density is offset by 4 km below the ring plane. The vertical range occupied by larger ring particles (larger than 600 nm) is about 10 times smaller than that for the nanograins.

ice fraction inferred for the C ring (6) than the two-third fraction that the CDA observed for nanograins. The difference is even larger for the A and B rings. The higher abundance of silicate grains seen in the nanograin ejecta population is a consequence of the ejecta production mechanism itself or a shorter lifetime of ice ejecta or a combination of both effects. The processes responsible for reducing the ice fraction in nanometer-sized ejecta as well as the latitudinal variation remain unclear.

Nature of the ring rain

Our simulations demonstrate that the CDA MA observed charged nanograins on their way from the surfaces of Saturn's main rings into the planet's atmosphere. The match between the MA rates and the H_3^+ latitudinal emission profile (Fig. 4B) is indicative of the CDA observing a ring rain composed of nanograins.

During their high-speed entry into Saturn's atmosphere, nanograins are expected to ablate and their material to be deposited as neutral atoms or molecules into the atmosphere. These additional neutrals reduce the local electron density, leading to a local accumulation of H_3^+. Models of the atmospheric capture of 10-nm ice particles predict their mass deposition near an altitude of 2000 km above the level of 1-bar pressure in Saturn's atmosphere (45). The deposited nanograin material would then diffuse downward and interact with the ionosphere, prolonging the lifetime of H_3^+ (46, 47).

We estimate from our simulation that the mass production rate for ejecta with sizes in the range of tens of nanometers, $\dot{\sigma}_{ej,nm}$, is 1,800 to 6,800 kg s^{-1} (38). Eighteen percent of these grains arrive at Saturn as ring rain, of which 30% (100 to 370 kg s^{-1}) are deposited in the mid-latitude region. This estimate is more than an order of magnitude higher than the 3 to 20 kg s^{-1} of water required to explain the observed H_3^+ emission (17). Nevertheless, given the uncertainties in the processes related to the ablation and mass deposition [e.g., (45)], our results agree reasonably well with the ring rain mass infall estimation, implying that the observed nanograin population is the cause of the ring rain effect (19, 21, 22).

The presence of the rings alters how the infall of exogenous material into the saturnian system affects the planet's atmosphere. The deposition of ring material through fast ejecta occurs predominantly at the equator and at southern latitudes and may also bear seasonal variations (38). Once deposited in the atmosphere, the nanograins as well as their ablation products likely serve as the nucleation seeds to form clouds and haze. The asymmetric mass deposition might contribute to the higher optical depth of the cloud or haze in Saturn's southern hemisphere (48–50).

Erosion of the rings

The gross erosion time of the rings ($t_g \sim \sigma_0/\dot{\sigma}_{ej}$), defined as the ratio of the rings' surface mass density (σ_0) to the ejecta flux ($\dot{\sigma}_{ej}$) (10), can be calculated on the basis of the nanograin mass production rate ($\dot{\sigma}_{ej,nm}$) derived from CDA measurements: $t_{g,nm} \sim 1 \times 10^8$ to 4×10^8 years
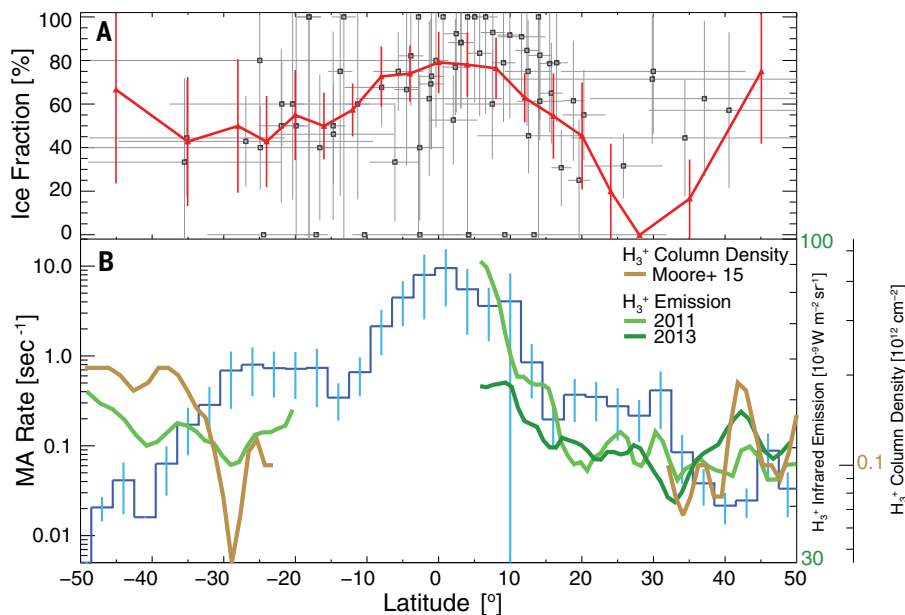
Fig. 4. Ice fraction and impact rate as a function of latitude. (**A**) The fraction of ice spectra exhibits a dependence on the latitude, with a maximum of 70 to 90% around the ring plane. Squares are dead-time–corrected ice fractions for 50 consecutive spectra; horizontal bars indicate the latitude range of the contributing spectra. The red line is derived directly from the spectrum number fraction without the dead-time correction. Error bars show the uncertainty associated with the counting statistics. (**B**) Comparison of the latitudinal profiles of the nanograin impact rate (blue histogram) and the H_3^+ infrared emission from Saturn's atmosphere observed between 2011 and 2013 (*18*) and the corresponding column density (*17*) (green and gold curves). The MA impact rate is a boxcar average weighted by the number of counts in each bin. The CDA measurements (ice fraction and impact rate) are shown as a function of the latitude of the spacecraft along its trajectory, whereas the H_3^+ profiles are shown as a function of planetocentric latitude on the planet. Away from the ring plane, the latitude at which nanograins enter the atmosphere is not the same as the latitude of the MA detection along the trajectory. This may explain the shift of pattern in comparing the MA rate and H_3^+ profiles.

(assuming a ring mass of 2×10^{19} kg). This is longer than the $t_g = 3 \times 10^5$ years derived from ejecta of all sizes (*10*), implying a mass fraction of 10^{-4} to 10^{-3} for ejecta in the size range of tens of nanometers.

Probing ring-planet interactions with nanodust

During the Cassini Grand Finale mission, the CDA MA collected material from Saturn's main rings. The gap between Saturn and the inner edge of the D ring is deficient in larger, micrometer-sized grains and populated predominantly by tiny ejecta particles tens of nanometers in radius. The grain dynamics are governed by Saturn's internal magnetic field, leading to a north-south asymmetric ring rain composed of nanograins into the planet's atmosphere. Our model calculations indicate that the nanograins are released predominantly from the B and C rings with speeds of a few hundred meters per second up to several kilometers per second. The observed fast population of tiny ejecta likely represents only a small fraction (<0.1%) of the total impact ejecta produced from the rings, and it is negligible regarding the angular momentum transport. However, the characteristic spatial distribution of the nanometer-sized grains resulting from the elec-tromagnetic dynamics makes them an element in the ring-planet interaction, and they are a probe of the ring composition. The high fraction of silicate nanograins in the CDA data, up to one-third of the identified mass spectra, likely does not provide a direct measure of ring composition but is related to ejecta production and the transport of eroding nanograins.

REFERENCES AND NOTES

1. G. P. Kuiper, D. P. Cruikshank, U. Fink, The composition of Saturn's rings. *Sky Telescope* **39**, 14 (1970).
2. C. B. Pilcher, C. R. Chapman, L. A. Lebofsky, H. H. Kieffer, Saturn's rings: Identification of water frost. *Science* **167**, 1372–1373 (1970). doi: 10.1126/science.167.3923.1372; pmid: 17778773
3. J. B. Pollack, A. Summers, B. Baldwin, Estimates of the size of the particles in the rings of saturn and their cosmogonic implications. *Icarus* **20**, 263–278 (1973). doi: 10.1016/0019-1035(73)90003-1
4. E. E. Epstein, M. A. Janssen, J. N. Cuzzi, Saturn's rings: 3-mm low-inclination observations and derived properties. *Icarus* **58**, 403–411 (1984). doi: 10.1016/0019-1035(84)90086-1
5. J. N. Cuzzi *et al.*, An evolving view of Saturn's dynamic rings. *Science* **327**, 1470–1475 (2010). doi: 10.1126/science.1179118; pmid: 20299586
6. Z. Zhang *et al.*, Cassini microwave observations provide clues to the origin of Saturn's C ring. *Icarus* **281**, 297–321 (2017). doi: 10.1016/j.icarus.2016.07.020
7. Z. Zhang *et al.*, Exposure age of Saturn's A and B rings, and the Cassini Division as suggested by their non-icy material content. *Icarus* **294**, 14–42 (2017). doi: 10.1016/j.icarus.2017.04.008
8. K. L. Thomas, G. E. Blanford, L. P. Keller, W. Klock, D. S. McKay, Carbon abundance and silicate mineralogy of anhydrous interplanetary dust particles. *Geochim. Cosmochim. Acta* **57**, 1551–1566 (1993). doi: 10.1016/0016-7037(93)90012-L; pmid: 11539451
9. R. H. Durisen, An instability in planetary rings due to ballistic transport. *Icarus* **115**, 66–85 (1995). doi: 10.1006/icar.1995.1079
10. J. N. Cuzzi, P. R. Estrada, Compositional evolution of Saturn's rings due to meteoroid bombardment. *Icarus* **132**, 1–35 (1998). doi: 10.1006/icar.1997.5863
11. S. Charnoz, L. Dones, L. W. Esposito, P. R. Estrada, M. M. Hedman, "Origin and evolution of Saturn's ring system," in *Saturn from Cassini-Huygens*, M. Dougherty, L. Esposito, S. Krimigis, Eds. (Springer, 2009), chap. 17, pp. 537–575.
12. J. Schmidt, K. Ohtsuki, N. Rappaport, H. Salo, F. Spahn, "Dynamics of Saturn's dense rings," in *Saturn from Cassini-Huygens*, M. Dougherty, L. Esposito, S. Krimigis, Eds. (Springer, 2009), chap. 14, pp. 413–458.
13. H. N. Latter, G. I. Ogilvie, M. Chupeau, The ballistic transport instability in Saturn's rings – I. Formalism and linear theory. *Mon. Not. R. Astron. Soc.* **427**, 2336–2348 (2012). doi: 10.1111/j.1365-2966.2012.22122.x
14. M. S. Tiscareno *et al.*, Observations of ejecta clouds produced by impacts onto Saturn's rings. *Science* **340**, 460–464 (2013). doi: 10.1126/science.1233524; pmid: 23620048
15. P. R. Estrada, R. H. Durisen, J. N. Cuzzi, D. A. Morgan, Combined structural and compositional evolution of planetary rings due to micrometeoroid impacts and ballistic transport. *Icarus* **252**, 415–439 (2015). doi: 10.1016/j.icarus.2015.02.005
16. J. O'Donoghue *et al.*, The domination of Saturn's low-latitude ionosphere by ring 'rain'. *Nature* **496**, 193–195 (2013). doi: 10.1038/nature12049; pmid: 23579676
17. L. Moore, J. O'Donoghue, I. Müller-Wodarg, M. Galand, M. Mendillo, Saturn ring rain: Model estimates of water influx into Saturn's atmosphere. *Icarus* **245**, 355–366 (2015). doi: 10.1016/j.icarus.2014.08.041
18. J. O'Donoghue *et al.*, Redetection of the ionospheric H_3^+ signature of Saturn's "ring rain". *Geophys. Res. Lett.* **44**, 11,762–11,769 (2017). doi: 10.1002/2017GL075932
19. T. G. Northrop, J. R. Hill, Stability of negatively charged dust grains in Saturn's ring plane. *J. Geophys. Res.* **87** (A8), 6045 (1982). doi: 10.1029/JA087iA08p06045
20. W.-H. Ip, On plasma transport in the vicinity of the rings of Saturn: A siphon flow mechanism. *J. Geophys. Res.* **88** (A2), 819 (1983). doi: 10.1029/JA088iA02p00819
21. J. E. P. Connerney, Magnetic connection for Saturn's rings and atmosphere. *Geophys. Res. Lett.* **13**, 773–776 (1986). doi: 10.1029/GL013i008p00773
22. T. G. Northrop, J. E. P. Connerney, A micrometeorite erosion model and the age of Saturn's rings. *Icarus* **70**, 124–137 (1987). doi: 10.1016/0019-1035(87)90079-0
23. W. Tseng, W. Ip, R. E. Johnson, T. A. Cassidy, M. K. Elrod, The structure and time variability of the ring atmosphere and ionosphere. *Icarus* **206**, 382–389 (2010). doi: 10.1016/j.icarus.2009.05.019
24. C.-M. Liu, W.-H. Ip, A new pathway of Saturnian ring-ionosphere coupling via charged nanograins. *Astrophys. J.* **786**, 34 (2014). doi: 10.1088/0004-637X/786/1/34
25. W.-H. Ip, C.-M. Liu, K.-C. Pan, Transport and electrodynamical coupling of nano-grains ejected from the Saturnian rings and their possible ionospheric signatures. *Icarus* **276**, 163–169 (2016). doi: 10.1016/j.icarus.2016.04.004
26. R. Srama *et al.*, The Cassini Cosmic Dust Analyzer. *Space Sci. Rev.* **114**, 465–518 (2004). doi: 10.1007/s11214-004-1435-z
27. S. Kempf *et al.*, High-velocity streams of dust originating from Saturn. *Nature* **433**, 289–291 (2005). doi: 10.1038/nature03218; pmid: 15662418
28. S. Kempf *et al.*, Composition of saturnian stream particles. *Science* **307**, 1274–1276 (2005). doi: 10.1126/science.1106218; pmid: 15731446
29. H.-W. Hsu *et al.*, Ongoing hydrothermal activities within Enceladus. *Nature* **519**, 207–210 (2015). doi: 10.1038/nature14262; pmid: 25762281
30. G. H. Jones *et al.*, Fine jet structure of electrically charged grains in Enceladus' plume. *Geophys. Res. Lett.* **36**, L16204 (2009). doi: 10.1029/2009GL038234
31. F. Postberg *et al.*, Discriminating contamination from particle components in spectra of Cassini's dust detector CDA. *Planet. Space Sci.* **57**, 1359–1374 (2009). doi: 10.1016/j.pss.2009.06.027
32. J. R. Göller, E. Grün, Calibration of the Galileo/Ulysses dust detectors with different projectile materials and at varying

impact angles. *Planet. Space Sci.* **37**, 1197–1206 (1989). doi: 10.1016/0032-0633(89)90014-7

33. J. N. Cuzzi *et al.*, HST-STIS spectra and the redness of Saturn's rings. *Icarus* **309**, 363–388 (2018). doi: 10.1016/j.icarus.2018.02.025

34. S. Kempf, Interpretation of high rate dust measurements with the Cassini dust detector CDA. *Planet. Space Sci.* **56**, 378–385 (2008). doi: 10.1016/j.pss.2007.11.022

35. D. A. Gurnett *et al.*, The Cassini radio and plasma wave investigation. *Space Sci. Rev.* **114**, 395–463 (2004). doi: 10.1007/s11214-004-1434-0

36. S.-Y. Ye *et al.*, Properties of dust particles near Saturn inferred from voltage pulses induced by dust impacts on Cassini spacecraft. *J. Geophys. Res. Space Phys.* **119**, 6294–6312 (2014). doi: 10.1002/2014JA020024

37. S.-Y. Ye *et al.*, Dust observations by the Radio and Plasma Wave Science instrument during Cassini's Grand Finale. *Geophys. Res. Lett.* 10.1029/2018GL078059 (2018).

38. Materials and methods are available as supplementary materials.

39. M. E. Burton, M. K. Dougherty, C. T. Russell, Saturn's internal planetary magnetic field. *Geophys. Res. Lett.* **37**, L24105 (2010). doi: 10.1029/2010GL045148

40. J. E. Wahlund *et al.*, In situ measurements of Saturn's ionosphere show that it is dynamic and interacts with the rings. *Science* **359**, 66–68 (2018). doi: 10.1126/science.aao4134; pmid: 29229651

41. M. Arakawa, M. Higa, Measurements of ejection velocities in collisional disruption of ice spheres. *Planet. Space Sci.* **44**, 901–908 (1996). doi: 10.1016/0032-0633(95)00119-0

42. Y. Shimaki, M. Arakawa, Experimental study on collisional disruption of highly porous icy bodies. *Icarus* **218**, 737–750 (2012). doi: 10.1016/j.icarus.2012.01.021

43. H. J. Melosh, Impact ejection, spallation, and the origin of meteorites. *Icarus* **59**, 234–260 (1984). doi: 10.1016/0019-1035(84)90026-5

44. M. Sachse, J. Schmidt, S. Kempf, F. Spahn, Correlation between speed and size for ejecta from hypervelocity impacts.
J. Geophys. Res. **120**, 1847–1858 (2015). doi: 10.1002/2015JE004844

45. O. Hamil, T. E. Cravens, N. L. Reedy, S. Sakai, Fate of ice grains in Saturn's ionosphere. *J. Geophys. Res.* **123**, 1429–1440 (2018). doi: 10.1002/2017JA024616

46. T. S. Stallard *et al.*, Peak emission altitude of Saturn's H_3^+ aurora. *Geophys. Res. Lett.* **39**, L15103 (2012). doi: 10.1029/2012GL052806

47. Y. H. Kim, J. L. Fox, J. H. Black, J. I. Moses, Hydrocarbon ions in the lower ionosphere of Saturn. *J. Geophys. Res.* **119**, 384–395 (2014). doi: 10.1002/2013JA019022

48. L. N. Fletcher *et al.*, Saturn's tropospheric composition and clouds from Cassini/VIMS 4.6–5.1μm nightside spectroscopy. *Icarus* **214**, 510–533 (2011). doi: 10.1016/j.icarus.2011.06.006

49. M. T. Roman, D. Banfield, P. J. Gierasch, Saturn's cloud structure inferred from Cassini ISS. *Icarus* **225**, 93–110 (2013). doi: 10.1016/j.icarus.2013.03.015

50. J. K. Barstow, P. G. J. Irwin, L. N. Fletcher, R. S. Giles, C. Merlet, Probing Saturn's tropospheric cloud with Cassini/VIMS. *Icarus* **271**, 400–417 (2016). doi: 10.1016/j.icarus.2016.01.013

51. D. T. Young *et al.*, Cassini plasma spectrometer investigation. *Space Sci. Rev.* **114**, 1–112 (2004). doi: 10.1007/s11214-004-1406-4

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIALS

RESEARCH ARTICLE SUMMARY

GAS GIANT PLANETS

Dust grains fall from Saturn's D-ring into its equatorial upper atmosphere

D. G. Mitchell*, M. E. Perry, D. C. Hamilton, J. H. Westlake, P. Kollmann, H. T. Smith,
J. F. Carbary, J. H. Waite Jr., R. Perryman, H.-W. Hsu, J.-E. Wahlund,
M. W. Morooka, L. Z. Hadid, A. M. Persoon, W. S. Kurth

INTRODUCTION: Ring material has long been thought to enter Saturn's atmosphere, modifying its atmospheric and ionospheric chemistry. This phenomenon, dubbed "ring rain," involves the transport of charged dust particles from the main rings along the planetary magnetic field.

RATIONALE: At the end of the Cassini mission, measurements by onboard instruments tested this hypothesis as well as whether ring material falls directly into the equatorial atmosphere. The final 22 orbits of the Cassini mission sent the spacecraft through the gap between the atmosphere and the innermost of the broad ring system, the D-ring.

RESULTS: The Magnetospheric Imaging Instrument—designed to measure energetic neutral atoms, ions, and electrons—recorded very small dust grains [8000 to 40,000 unified atomic mass units (u), or roughly 1- to 3-nm radius] in two sensors. At 3000-km altitude, a peak rate of ~300,000 counts s^{-1} was detected by one sensor as Cassini crossed the equatorial plane. At lower altitude (1700 to 2000 km), a second sensor recorded positively charged dust in the upper atmosphere and ionosphere over a size range of ~8000 to 40,000 u (~1 to 2 nm, assuming the density of ice).

Consistent with this observation, larger dust in the 0.1- to 1-μm range was detected by the Cassini Dust Analyzer and the Radio and Plasma Wave Science instrument.

CONCLUSION: We modeled the interaction of dust with the H and H_2 exospheric populations known to populate the gap. Collisions between small dust grains and H atoms provide sufficient drag to de-orbit the dust, causing it to plunge into the atmosphere over ~4 hours. The analysis indicates that at least ~5 kg s^{-1} of dust is continuously precipitating into the atmosphere. At 3000-km altitude, the dust is distributed symmetrically about the equator, mostly between ±2° latitude with a peak density of ~0.1 cm^{-3}. On the wings of the distribution, consistent with ring rain transport along the magnetic field, almost all of the dust was observed to be charged. At the 2000 to 1700 km altitude, the dust has reached a diffusive terminal velocity and, although showing some bias toward the equator, is ordered mostly by a scale height of ~180 km in altitude. The most probable source for this dust population is the innermost bright ringlet of the D-ring, known as the D68 ringlet.

We predict that this kinetic process generates a highly anisotropic neutral hydrogen population, concentrated near the equatorial plane with periapses between ~4000 and 7000 km, and apoapses ranging to as high as 10 Saturn radii, with a small fraction on escape trajectories. ∎
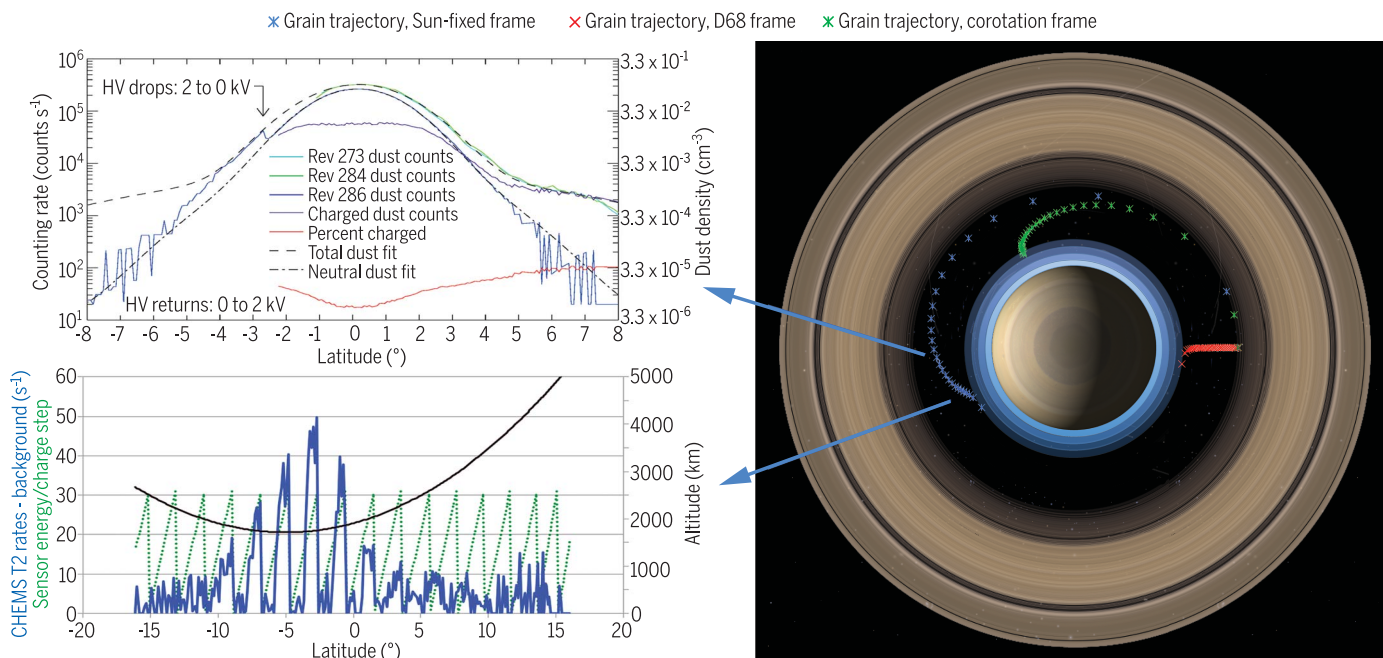
Ring dust. (Top left) Data and model fits for the equatorial dust population near 3000-km altitude for three orbits through the D-ring gap. HV, plate detector high voltage. Red line uses left scale (percent). (Bottom left) Dust counts (blue) from ~2000 to 1700 km (modulated by sensor energy/charge steps, green), ordered by altitude and latitude, consistent with diffusive transport. (Right) Model trajectory of a dust particle in three frames of reference, as collisions with exospheric hydrogen degrade its velocity. Saturn has been shrunk to expand the gap for clarity.

RESEARCH ARTICLE

GAS GIANT PLANETS

Dust grains fall from Saturn's D-ring into its equatorial upper atmosphere

D. G. Mitchell[1]*, M. E. Perry[1], D. C. Hamilton[2], J. H. Westlake[1], P. Kollmann[1], H. T. Smith[1], J. F. Carbary[1], J. H. Waite Jr.[3,4], R. Perryman[3], H.-W. Hsu[5], J.-E. Wahlund[6], M. W. Morooka[6], L. Z. Hadid[6], A. M. Persoon[7], W. S. Kurth[7]

The sizes of Saturn's ring particles range from meters (boulders) to nanometers (dust). Determination of the rings' ages depends on loss processes, including the transport of dust into Saturn's atmosphere. During the Grand Finale orbits of the Cassini spacecraft, its instruments measured tiny dust grains that compose the innermost D-ring of Saturn. The nanometer-sized dust experiences collisions with exospheric (upper atmosphere) hydrogen and molecular hydrogen, which forces it to fall from the ring into the ionosphere and lower atmosphere. We used the Magnetospheric Imaging Instrument to detect and characterize this dust transport and also found that diffusion dominates above and near the altitude of peak ionospheric density. This mechanism results in a mass deposition into the equatorial atmosphere of ~5 kilograms per second, constraining the age of the D-ring.

Particle influx from Saturn's rings to its upper atmosphere, termed "ring rain," was invoked to account for ionospheric electron densities measured by the Voyager spacecraft, which were lower than expected from solar extreme ultraviolet ionization (1, 2). Ring rain is expected to occur at altitudes below ~0.5 R_S [where 1 R_S (Saturn radius) = 60,268 km]. In this region, magnetic drag on charged grains can slow them below local Keplerian orbital velocity, and the gravitational force dominates over the electromagnetic and centrifugal forces that otherwise keep such small (~100-nm- to 1-μm-sized) charged dust grains in orbit. The charged grains are then pulled along magnetic field lines into the planet's atmosphere (3–6). The water-based ice grains in the atmosphere absorb electrons, thereby reducing the ionospheric electron density, primarily at latitudes that lie at the atmospheric feet of magnetic field lines connected to the rings (2). This reduction has been remotely observed from ionospheric occultation measurements (7, 8) and from ground-based telescope observations (6). Simulations of ring rain suggest that charged dust precipitation should be greatest at the magnetic foot-points of the main rings (>15° from the equator), whereas the ionospheric

electron densities are observed to be lowest near the equator (8–10). The ring rain mechanism does not predict this, because the equator is both close to the region of highest ionizing solar flux and inaccessible to the charged dust that would follow the magnetic field.

Dust particles have previously been detected with Cassini instruments—including the Low Energy Magnetospheric Measurements System (LEMMS) instrument, the Cassini Plasma Spectrometer (CAPS) electron spectrometer (ELS) and ion mass spectrometer (IMS), and the Radio and Plasma Waves Spectrometer (RPWS)—which all detected dust in the plumes emanating from the moon Enceladus (11–13). The ELS also detected large negative molecules in the ionosphere of large moon Titan (14).

We report in situ measurements of ring-material precipitation measured during the final, proximal orbits of the Cassini spacecraft. Referred to by rev (orbit) number, these orbits took the spacecraft very close to the planet within the D-ring. Of the 293 revs during the Cassini mission, the final 23 were the proximal orbits. Our measurements were made during revs 273, 284, 286, and 288 through 293 with two subsystems of the Magnetospheric Imaging Instrument (MIMI) (15), namely the ion and neutral camera (INCA) and the charge energy mass spectrometer (CHEMS).

Dust measurements in the gap between Saturn and its D-ring

Only three of the proximal orbit equator crossings were useful for INCA measurements, because the dust ram velocity vector (the vector sum of prograde circular Keplerian orbital motion and the spacecraft velocity) must be within INCA's 90°-by-120° field of view (FOV). The INCA obser-

vations during these three orbits—at 2850-, 3000-, and 3100-km altitude—showed distinct signals, which we attribute to dust (Fig. 1). INCA records particles both when they enter the detector at an initial microchannel plate (MCP, a start count) and when they strike a final MCP (a stop count) (16). The dust signals are inconsistent with known backgrounds such as penetrating electrons or sunlight, both of which produce approximately proportional responses on the INCA start and stop MCP detectors, whereas the dust signal is exclusive to the INCA start MCP. All three of these equator crossings also had similar (though not identical) features on the outbound orbital egress. For revs 273 and 284, the count rates dropped much more quickly on the egress than they rose on the ingress, owing to a rotation of Cassini about the spacecraft's high-gain antenna (z) axis, which is a planned maneuver used to facilitate calibration of the magnetometer for other studies (17). Just after the peak counting rates were reached (at the equator), the spacecraft rotation took the dust ram vector out of INCA's FOV, so the rate dropped quickly as the dust stream no longer entered the instrument. On rev 286, the response was much more symmetric because Cassini maintained a nearly fixed attitude relative to the dust ram direction, but there is a jump in the rate curve on the egress, where the start rate abruptly nearly doubled before continuing its decline. This jump was produced by a sudden drop in the INCA charged-particle deflection plates voltage from 2 to 0 kV, which was caused by the immersion of the spacecraft in ionospheric density high enough to short-circuit the deflection plates as Cassini approached its minimum altitude during this rev. To exclude ions when INCA is in its energetic neutral atom imaging mode, these deflection plates were biased at 2 kV for rev 286 but were held at 0 V (ion mode) during revs 273 and 284 (16).

In this near-equatorial region containing Saturn's inner radiation belt, the counting rates on INCA's stop MCP are driven primarily by penetrating particles. The inner radiation belt (18) is composed primarily of very-high-energy protons (≥300 MeV) that easily penetrate the walls of the INCA sensor and produce counts in the detector proportional to both the start and stop areas. The correspondence between the stop rates and the fraction of the start rates that are driven by penetrators can be used to predict and then subtract the contribution from penetrators to the start rates, producing a background-corrected version of the start rates that we attribute to dust impacts (16).

Figure 2 shows a plot of dust count rate as a function of spacecraft planetocentric latitude for the three different equator crossings. The rates have been background corrected (16) but otherwise are the raw rates. The count rate curves for revs 273 and 284 are almost identical, despite being taken more than 2 months apart. From this, we infer that the dust population is constant over this time and altitude scale. For revs 273 and 284, no voltage was applied across the INCA deflection plates, so the dust is counted independent of its charge state. We performed a trial-and-error fit to the derive a count rate

¹Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA. ²Department of Physics, University of Maryland, College Park, MD, USA. ³Southwest Research Institute, Division of Space Science and Engineering, San Antonio, TX, USA. ⁴The University of Texas San Antonio, San Antonio, TX, USA. ⁵Laboratory for Atmospheric and Space Physics, University of Colorado, Boulder, CO, USA. ⁶Swedish Institute of Space Science, Uppsala, Sweden. ⁷Department of Physics and Astronomy, University of Iowa, Iowa City, IA, USA.
*Corresponding author. Email: donald.g.mitchell@jhuapl.edu

function (*CRF*) for these dust count profiles, using the sum of two Gaussian distributions:

$$CRF_{TotalDust} = A\exp\{\tfrac{1}{2}[(\lambda - 0.17°)/\sigma_A]^2\} + B\exp\{\tfrac{1}{2}[(\lambda - 0.17°)/\sigma_B]^2\} \quad (1)$$

We found that the constant that scales the peak amplitude of the equatorial counting rate (*A*) = 316,000, the constant that scales the amplitude of the higher latitude wings of the distribution (*B*) = 6000, the Gaussian width of the equatorial peak (σ_A) = 1.4°, and the Gaussian width of the higher latitude component (σ_B) = 5°; λ is the planetographic latitude. The small 0.17° offset may derive from the known northward offset of the magnetic equator from the spin equator of Saturn, although it is formally only a parameter needed for a good fit. Uncertainties in *A* and *B* are ~5 and ~10%, respectively; uncertainties on σ_A and σ_B are ~5%, and the uncertainty on λ is ~±0.2°.

For rev 286, the voltage across INCA deflection plates was 2 kV. This setting removed all charged dust up to about 80 or 90 keV per elementary charge (keV/e) and most of the charged dust up to roughly 200 keV/e. Given the combined velocity of the spacecraft and the dust, the impacts on the INCA start foil are at about 31 km s^{-1}. By using this velocity, 200 keV/e is equivalent to a particle mass of just under 40,000 unified atomic mass units (u). Because the particle velocity is the same for all particles dominated by the spacecraft velocity, particle mass scales linearly with energy. The ingress for the rev 286 distribution was also fitted by using two Gaussian components (Fig. 2), as below:

$$CRF_{NeutralDust} = C\exp\{\tfrac{1}{2}[(\lambda - 0.17°)/\sigma_C]^2\} + D\exp\{\tfrac{1}{2}[(\lambda - 0.17°)/\sigma_D]^2\} \quad (2)$$

where *C* = 250,000, *D* = 1550, σ_C = 1.22°, and σ_D = 2.15°. The definition and uncertainties for these parameters are similar to those for Eq. 1.

During the egress on rev 286 toward south latitudes, Cassini was moving to lower altitudes (the minimum altitude was reached at about −5° latitude). When the ionospheric density short-circuited INCA's deflection plates, the voltage dropped suddenly from 2 to 0 kV and resulted in the entry of the charged dust that had been excluded by the deflection system. This accounts for the abrupt increase in the dust counting rates by about a factor of two, as seen in Fig. 2. The resulting count rate jumped up to the total dust fit curve, confirming the constancy of the dust population for all three revs and giving us additional confidence in our approach to estimating the fraction of the population that is both charged and below the INCA deflection system maximum sensor energy/charge (*E/q*) limit. The INCA deflection system removed charged dust independent of its sign—both negative and positive dust are removed with equal effectiveness. Therefore, we cannot determine from these data what proportion of the charged dust is positive or negative, only the fraction that is charged and has *E/q* ≤ ~200 keV/e, or, for this velocity, ~40,000 u/e. Previous work has shown that, in this size range,

grains will acquire only three likely charge states, −1 *e*, 0, or +1 *e*, with the most probable charge state at higher altitudes being neutral [e.g., (*13*, *19*–*21*)], and the charging times are hours to days (*19*, *21*, *22*).

Near the equatorial peak, only 20% of the dust measured is charged, but that fraction rises to 100% by ±6° latitude. This can be attributed to different dynamics for the neutral and charged components. Charged dust in this size range is constrained by the magnetic field, so it cannot fall radially downward but instead moves to lower altitude along the magnetic field (*1*–*5*). This

process shifts an equatorial dust source to off-equatorial latitudes, which we invoke to explain the 100% charged-dust population detected in the higher-latitude wings of the total dust distribution. The absence of neutral dust in this higher-latitude population reflects both that this charged population is all less than 40,000 u (otherwise the INCA deflection system would not have removed it in rev 286) and that at this altitude, the drag-related transport time for a grain from the D-ring to the 3000-km altitude is much shorter than the time for a grain to change charge state. Because this charged-dust transport process
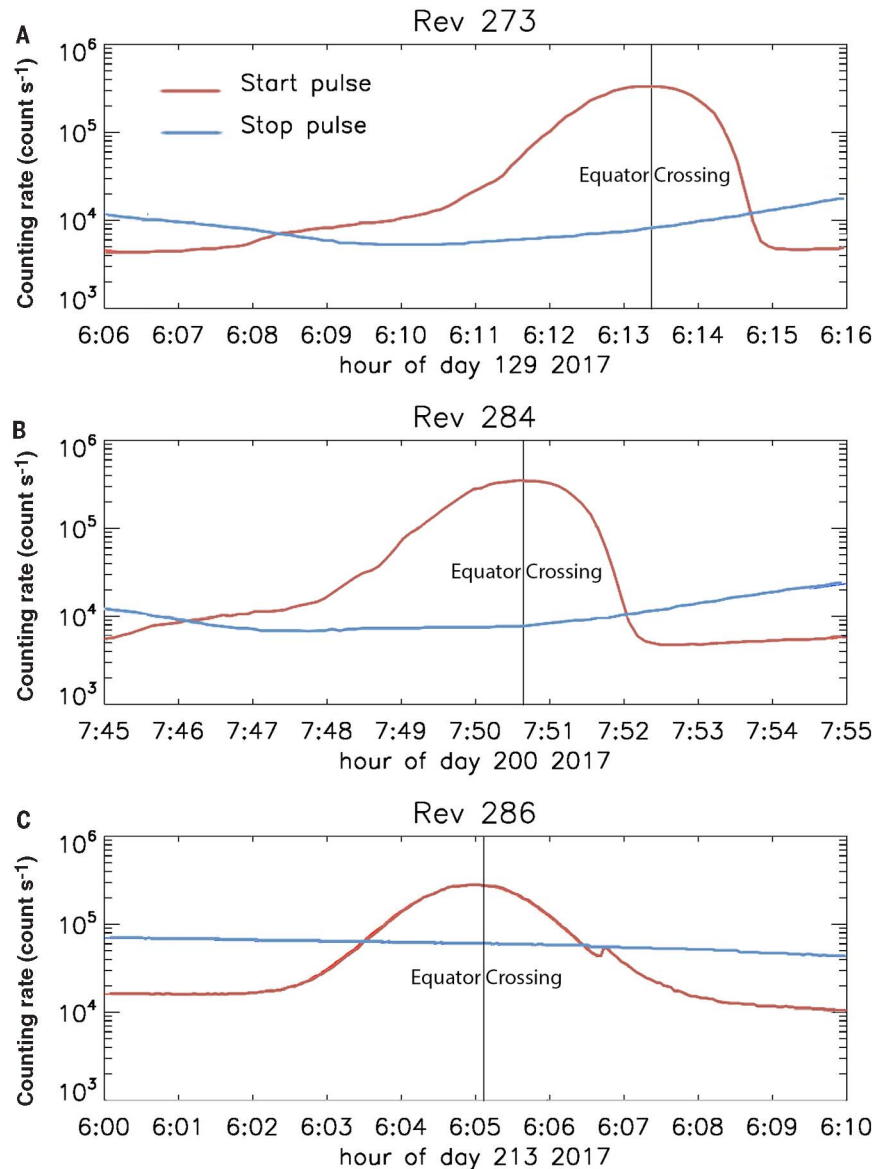

Fig. 1. INCA detector counting rates. (**A** to **C**) Raw INCA start and stop counting rates for Cassini equator crossings on revs 273 (A), 284 (B), and 286 (C). As Cassini traverses through the equatorial region during the proximal orbits, the INCA start rates climb, owing to increased dust near the ring plane. Counting rates on the start and stop plates for these three proximal-orbit periapses that had the ram vector in the INCA FOV show evidence of dust, evident as enhancements in start rates (red traces) without changes in the stop rates (blue traces) as Cassini crossed the equator. In each instance, the start rate reaches its maximum at the equator crossing time.

Fig. 2. Dust counting rates and analysis. Dust counting rates (left scale) for revs 273 (light cyan), 284 (green), and 286 (blue) after background subtraction, plotted against Saturn latitude. The spacecraft approached the equator from the north, so time runs from right to left in this representation. The rates for revs 273 and 284 have been truncated at the point where the dust ram vector begins to leave the INCA FOV. Also shown are model fits to the rev 273 and 284 data (dashed line, total dust fit); the fitted ingress and peak of the rev 286 data (dashed-dotted line, neutral dust fit); and curves derived from those models. The charged dust counts (purple) curve repre-sents the difference between

the total dust fit and the rev 286 measured dust flux, which we take to be due to dust that is both charged and in the E/q range of less than 180 to 200 keV/e. The red curve at the bottom represents the percent (also read on the left scale) of the total dust contributed by the charged dust counts. All of the curves (with the exception of the red one) can also be read from the right-hand scale, which converts them to the measured dust density (see text). The label "HV drops" indicates that the deflection plate high voltage (HV) dropped from 2 to 0 kV at this point; "HV returns" indicates that over this latitude location, the deflection plate HV recovered from 0 to 2 kV. cts, counts.
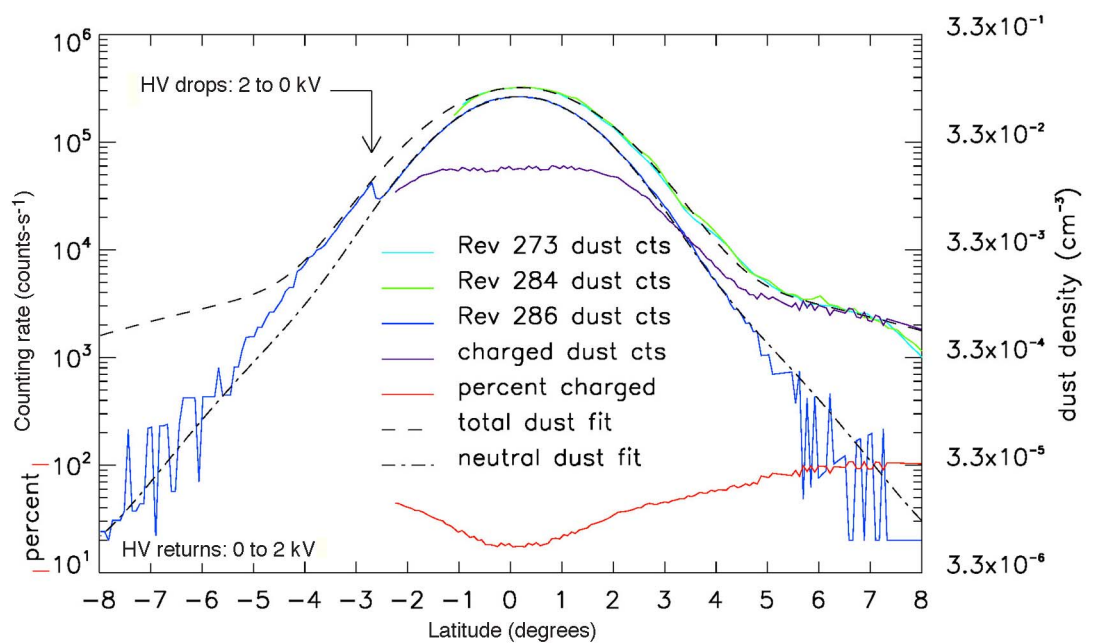
has been studied previously (*3*, *4*) and this higher-latitude component forms ~1% of the equatorial peak, we hereafter focus on the neutral compo-nent, which feels no electromagnetic forces, and thus its transport is dominated by gravity and collisions with the atmosphere.

Figure 2 includes a conversion of the count rate into density. The primary assumption in calculat-ing the density is that the efficiency for dust detection by INCA is 100% above some minimum energy that can penetrate the INCA start foil. The foil has an areal mass density of 10 μg cm^{-2}, so we estimate this energy to be about 50 to 60 keV. At the dust ram velocity, this is probably a valid assumption, because the penetration of the foil and the generation of a plasma by the impact of the dust with the foil is almost unavoidable. Under this assumption, the density (N) is simply

$$N_{\text{Dust}} = CR_{\text{Dust}}/(V_{\text{DustRam}}A_{\text{Foil}}) \qquad (3)$$

where CR_{Dust} is the INCA start counting rate after background subtraction, given by Eqs. 1 or 2; V_{DustRam} is the dust ram velocity (~31.5 km s^{-1}); and A_{Foil} is the start foil area, corrected for the angle between the ram velocity and the normal to the foil and corrected for the transmission of the foil support grid. The foil area is 2 cm^2, the angle from the foil normal ranges from about 50° to 55° for the attitude configurations near the equator crossings, and the grid transmission is 82%. The corrected foil area then ranges be-tween 1.05 and 0.94 cm^2, so we adopt 1.0 cm^2 for this calculation. The peak density at the equa-tor is then $N_{\text{Dust}} = 0.1$ cm^{-3} for all three revs, de-spite the fact that Cassini's equatorial altitude

varied between 2850 and 3100 km. As we show below from dust flux–continuity considerations, this is the altitude range where the transport transitions from nearly free falling to collision dominated—where the vertical velocity becomes constant and density reaches a local minimum—so density remains approximately constant at these altitudes.

Dust-grain trajectories

In Fig. 3, A and B, we show the angular distri-bution of the dust from the two equator crossings during which Cassini was rolling about the spacecraft z axis. The phase of the roll angle was identical for both revs 273 and 284 to optimize ionosphere and neutral atmosphere observations by the Ion Neutral Mass Spectrometer (INMS) (*23*) at the lowest altitude in the orbit, which was reached at about −5° latitude on each rev. Thus, the position of the INCA FOV varied with latitude in the same way over both revs. A comparison of the measured counting rates with the extrapolated fit to the ingress and peak reveals a close match for both revs until about a minute past the peak, when the signal begins to deviate from the fit. As the angular distribution of the dust that is ramming into the instrument (from the combi-nation of spacecraft and dust orbital motion) begins to cross outside the edge of the INCA FOV, the counting rates drop quickly. The ratio be-tween the data and the fit (F_{10}, where $F_{10} = 10 \times CR/CRF$) drops from ~1 to ~0 as the dust leaves the INCA FOV (Fig. 3, A and B). A similar signa-ture appears at the onset of the signal, about a minute after the start of each plot, but because the dust density is much lower, the statistics on

the rise are less well characterized and more subject to background subtraction uncertainties. The angle $\alpha_{\text{Keplerian}}$ that the edge of the INCA FOV makes with the calculated ram vector correspond-ing to the local circular Keplerian motion is also plotted in Fig. 3, A and B. If the dust velocity was centered on the value corresponding to circular Keplerian motion, then the point at which the ratio of the measured dust signal dropped to 0.5 relative to the fit would be centered at $\alpha_{\text{Keplerian}} = 0$. Instead, the ratio drops to ~0 at $\alpha_{\text{Keplerian}} = 0$. To determine the angular distribution of the dust ram directions, we differentiate the curve describing the cutoff with rotation, F_{10}. To do this, we plot F_{10} against $\alpha_{\text{Keplerian}}$ and fit a poly-nomial to the resulting points (Fig. 3C). We then differentiate the polynomial and plot that func-tion (normalized) against $\alpha_{\text{Keplerian}}$ (Fig. 3D). The resulting ram angle distribution has a peak at about 5°, dropping to near zero at 0° and ~14°. The (azimuthal) velocities that would result in this distribution range from 100% Keplerian at 0° to 85% Keplerian at 5° to 72% Keplerian at 14° (the distribution is not symmetric about its peak).

At this altitude, any neutral dust particle with velocity less than about 98.5% local circular Kep-lerian (virtually all of the neutral particles mea-sured by INCA) has an orbital periapsis below the top of Saturn's atmosphere (defined as the height at which the atmospheric pressure is 1 bar). Regardless of its subsequent collision history, any such dust particle is destined to precipitate into the atmosphere. Any additional collisions expe-rienced on the way down simply determine where, not whether, it will precipitate. We consider the effect of collisions below.
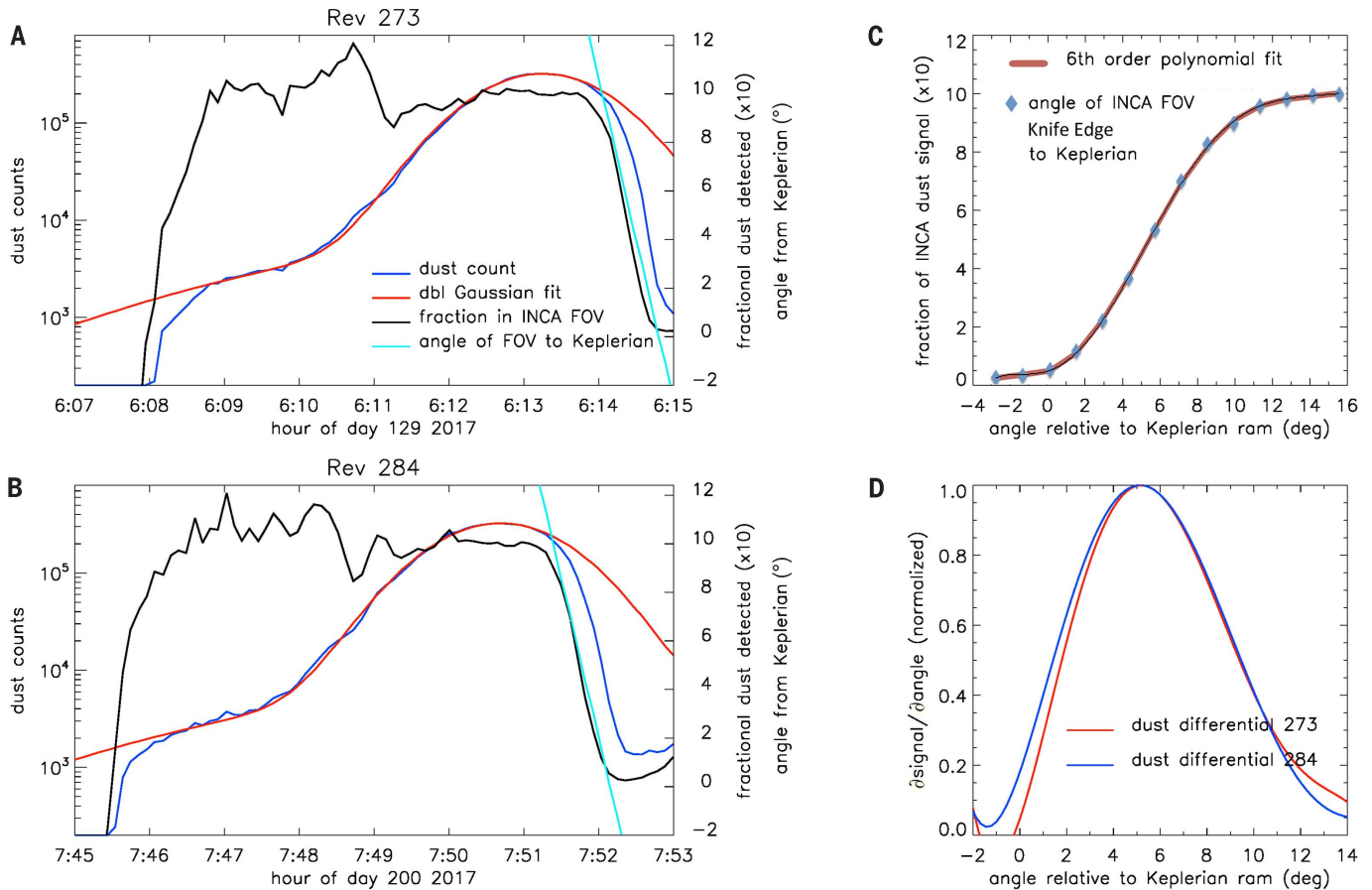
Fig. 3. Determination of dust angular distribution. (**A** and **B**) For rev 273 (A) and rev 284 (B), the red curve is a double (dbl)-Gaussian function fit to the dust count rate data (left scale); the blue curve is the dust count rate measured by INCA (left scale); the black curve is the ratio of the INCA dust count rates to the fit, multiplied by 10 ($F_{10} = 10 \times CR/CRF$, right scale); and the cyan curve is the angle between the sharp edge of the INCA FOV and the calculated ram vector, assuming circular Keplerian dust ($\alpha_{Keplerian}$), as the spacecraft rotates about its z axis. (**C**) The blue points show F_{10} (for rev 284; rev 273 is essentially identical) plotted against $\alpha_{Keplerian}$; the red curve is a polynomial fit to F_{10} versus $\alpha_{Keplerian}$. (**D**) The (normalized) differentiation of the polynomial fit function is shown, where rev 273 is in red and rev 284 is in blue. This represents the angular distribution of the dust ram response relative to $\alpha_{Keplerian}$, showing that rather than being centered on $\alpha_{Keplerian} = 0$, as would be expected for dust in circular Keplerian motion, the dust ram is centered 5° away from zero.

Entry of dust into Saturn's atmosphere

We have modeled the trajectories of (uncharged) Keplerian dust grains starting just above the altitude of the D68 ringlet in Saturn's D-ring, as they collide with atoms (primarily H at the highest altitude, with H_2 and He becoming important below 4000 to 5000 km). We track their changes in velocity due to the collisions and the effects of gravity. Figure 4 shows the results of this calculation for a specific dust grain size, 19,300 u, which corresponds to ~100 keV at the spacecraft ram velocity. The collisions with H, H_2, and He are shown, based on an atmospheric model (24), with H_2 densities adjusted to conform with in situ measurements by INMS (25). We sum the collisions with H, H_2, and He, weighted by their atomic masses, by modeling collisions with a species of mass 2 u for H_2 and scaling the number of collisions with H and He by one-half and two, respectively. The horizontal component of the velocity is referenced to the atmospheric corotation velocity—that is, if the velocity drops to zero, the grain is no longer moving with respect to the rest frame

of the atmosphere. The vertical velocity peaks at about 2700-km altitude, below which collisions modify its velocity profile until about 1700 km (depending on particle radius) where the particle has reached its terminal velocity, and its transport becomes more fully diffusive.

From this calculation, we also determine the vertical velocity for the dust as a function of altitude. The vertical velocity peaks just below 3000-km altitude at about 2.5 km s^{-1}. By using this velocity, a dust density of 0.1 cm^{-3}, assuming an average grain size of 15,000 u (~2.5×10^{-23} kg) for the INCA measurements and integrating over an area between ±1° latitude and over the circumference of Saturn at 3000-km altitude, we find $(2.5 \times 10^{-23}$ kg$) \times (0.1$ cm$^{-3}) \times 2\pi \times (63,280 \times 10^5$ cm$) \times 2° \times (\pi/180°) \times (63,280 \times 10^5$ cm$) \times (2.5 \times 10^5$ cm s$^{-1}) =$ ~5.5 kg s^{-1} mass flux into the atmosphere. This is the flux for the particles that INCA can measure, that is, those larger than approximately 10,000 u. As described above, the upper limit is poorly constrained but is likely less than 40,000 u. Both the assumption of the average grain size and

the ~10% uncertainty in average vertical velocity contribute to uncertainties as high as 50% in the mass flux estimate. Given the likelihood that the size distribution of the dust is probably a power law with a slope of at least 3 and possibly as steep as 6 (26), the total mass flux of all dust particles is probably much higher.

This mass flux (~1.7×10^{-13} g m^{-2} s^{-1}) is about an order of magnitude higher than that estimated for 0.5- to 100-μm interplanetary dust into Saturn's atmosphere (27). However, there are some important differences. First, this is D-ring dust, and this rate of influx may be a relatively recent development, beginning in about 2014 with a disruption event in the D68 ringlet (25, 26). Second, this D-ring dust comprises much smaller grains than those previously modeled (27). Third, the dust velocity relative to the atmosphere at Saturn is much lower than for interplanetary dust, either for Saturn or for Earth, because it is falling from a nearby Keplerian orbit rather than from infinity. The average neutral-atmosphere atomic mass is much smaller at Saturn than at Earth. This,
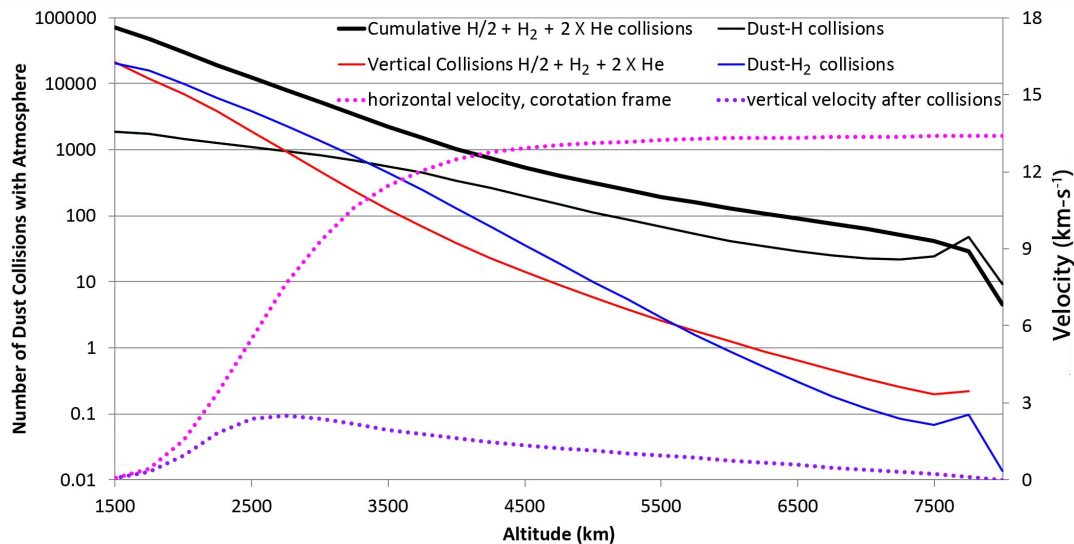
Fig. 4. Collisions of dust grains with atmospheric atoms and resulting velocities. Cumulative number of collisions (left scale, solid curves) between a dust grain (19,300 u and 1.8-nm radius, in this example) and Saturn's atmosphere, as a function of altitude above Saturn's 1-bar level. At higher altitudes, collisions with H (thin, black line) dominate the horizontal collisions. Below ~3300 km, H_2 collisions (blue line) dominate. The two dotted curves show the velocity components (horizontal and vertical, right axis scale) of the grain as it is slowed by collisions with atmospheric atoms. Vertical collisions (red line) become important below ~2800 km, where the vertical velocity profile begins to plateau.

together with the lower relative velocity, means that these particles do not ablate as interplanetary dust does in Earth's upper atmosphere and so do not break up or directly affect the chemistry of the atmosphere in the same way as interplanetary dust does in Earth's mesosphere. Finally, this influx is limited to the equatorial region [although charged dust flux from the main rings also likely exceeds the interplanetary dust flux (28)].

The distribution in latitude measured by INCA arises directly from the angular scattering caused by the collisions that degrade the grain velocities. The grains—which we argue below begin their descent from the inner edge of the D-ring, most likely from the D68 ringlet [see, for example, (29)]—are initially confined very close to the equator, likely within a few kilometers or less. The angular scattering due to H atoms during their deceleration will result in a random walk relative to the equatorial source, gradually spreading the distribution as it descends.

Figure 5 shows the model dust density profile (for the size range to which INCA is sensitive) as a function of altitude, normalized to the INCA measurement of 0.1 cm^{-3} at 3000 km. The continuity calculation does not include latitudinal diffusion, so the predicted densities below the altitude of the INCA measurements are, in that sense, upper bounds. Within the grain size range measured by INCA, the predicted density dependence can be modeled using a scale-height formalism at the lower altitudes (see below), with a smaller grain size distribution resulting in higher densities at the lower altitudes. Because the dust grain size is only loosely constrained by the INCA measurements, the scaling by the observed number density of 0.1 cm^{-3} at 3000-km altitude may represent a size range that is larger than that appropriate to the MIMI CHEMS observations (below) and much larger than the molecular

grains (25). Because we have no direct measurements of neutral H, the model values for H are not well constrained. In Figs. 4 and 5, we used the higher-density model for H (24), which has a considerable effect on the collision history. Because by the 3000-km altitude the collision dynamics are dominated by H_2, reducing the H to much lower (0.1%) than the predicted values (24) renders the impact of collisions with H negligible, yet only alters these velocity profiles at and below 3000 km slightly, with the horizontal velocity at 3000 km changing from 9.3 km s^{-1} in the atmosphere frame to 10.2 km s^{-1}, a ~10% difference. Thus, the observations do not strongly constrain the H density. Such a reduction in the H density would have an effect on the higher-altitude dust density profile predicted by the model (Fig. 5), leading to a stronger upward gradient in dust density and higher dust densities above 3000 km than the model currently indicates. Because we have no direct measurements of the equatorial dust density above 3000 km, we cannot presently constrain the H density with much precision.

Influence on Saturn's atmosphere

This influx of dust grains and their collisions with the very high atmosphere has implications for the atmospheric structure both below and above the altitudes sampled by Cassini. The effects on the ionosphere are substantial (25), and there are also effects on the high-altitude exosphere. In slowing from Keplerian speeds to corotational speeds, each dust grain undergoes thousands of kinetic collisions with H atoms and H_2 molecules at altitudes where these constituents' mean free paths are very long. Each such (assumed elastic) collision, if effective in slowing the grain, is also very effective in speeding up the colliding atom or molecule. For example, a head-on

collision between a Keplerian (velocity ~24 km s^{-1}) grain of 10,000 u and a H atom moving at the 10–km s^{-1} corotational speed in the same direction (plus some small thermal velocity) leaves the grain moving at 23.997 km s^{-1} and the atom moving at 38 km s^{-1}. The latter is above the escape velocity for Saturn. This implies that this process will produce perhaps thousands (at least hundreds) of high-velocity H atoms for every dust grain that precipitates into the atmosphere. However, these atoms do not have an isotropic distribution. Because the dust is confined to the equator, the collisions which impart the greatest velocity are (nearly) head-on, and the dust motion is mostly horizontal, the atoms that receive high speed boosts will be moving nearly tangentially with the top of the atmosphere and so are likely to continue to very high altitudes without additional collisions. Their trajectories are expected to be roughly parallel to the equator, so they should constitute a quasi-equatorial population of neutral hydrogen with periapsis at the top of the atmosphere and apoapsis at several R_S. Some H_2 may also achieve similar trajectories, although the probabilities are lower because of the lower H_2 densities at the highest altitudes where escape without collision is most likely.

This energetic, anisotropic H and H_2 population is very unlikely to be lost to collision with the rings, which constitute a very thin target for atoms or molecules launched from the equator. We suggest that this is the mechanism behind the equatorial hydrogen plume detected by the Cassini Ultraviolet Imaging Spectrograph (UVIS) instrument (21).

Dust at lower altitudes

Whereas the ram vector entered INCA's aperture on the revs discussed above, for several later revs at lower near-equatorial altitudes, the dust ram

Fig. 5. Predicted altitude profile of dust density based on dust flux continuity. The blue line uses the continuity of the dust grain flux determined from the calculated vertical velocities combined with the dust density (in the size range measured by INCA) at ~3000 km (0.1 cm^{-3}) to predict the densities above and below the altitude of the INCA measurements. The blue diamond shows the altitude of the D68 ringlet, which we suggest is the primary source for the dust observed by Cassini. The upturn at low altitude depends on the dust radius and material density. Atmospheric mass density is plotted in red on the right scale. The decreasing dust density from high to low altitude reflects the increasing vertical velocity as the particles deorbit, until, at ~2800 km,

the atmospheric mass density is sufficient to begin decelerating the falling particles, transitioning the transport to diffusive. For the 19,300-u particle modeled, the vertical transport is dominated by collisions below ~2200 km, leading to the density being well characterized by a scale height at those altitudes, as was observed (close fit to the dotted green curve, based on a 180-km scale height).
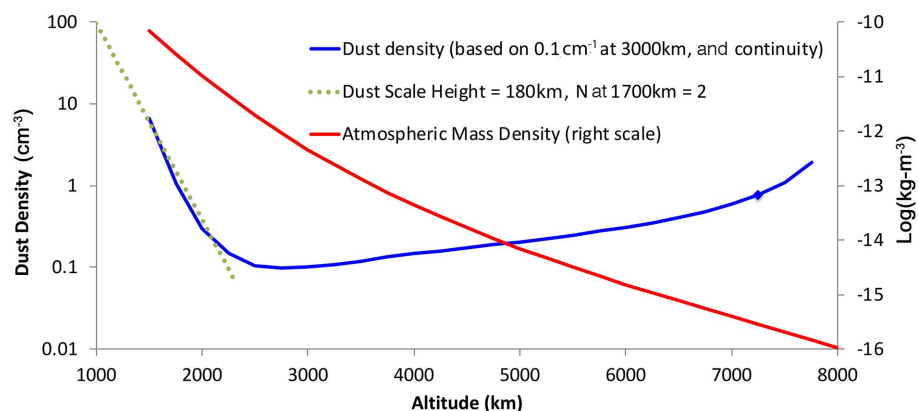
Fig. 6. CHEMS dust counting rates for Rev 288.
(**A**) Raw start counts per second from the CHEMS T1, T2, and T3 telescopes as a function of latitude for the periapsis pass on rev 288 (time runs from right to left). The strongest signal is in T2, which was oriented to detect positively charged dust grains in the combined spacecraft and corotation vector ram flow. The square wave pattern is created as the CHEMS ESA steps through its E/q selection voltages—grains selected in the lower end of the CHEMS E/q range do not have sufficient energy to penetrate the start foil and generate start electrons. On this orbit, the inner radiation belt (18) creates background counts from high-energy penetrating charged particles, especially at latitudes beyond about 10° from either side of the equator. (**B**) Using the T3 start counts to represent background (bkgd) from penetrating particles and scaling the T3 rates to the same level as the observed backgrounds in T2, we subtract this background from the T2 rates, leaving the dust counts as signal. Also plotted are the spacecraft altitude (black, right axis) and the corresponding ESA HV step for each measurement (green, left axis), with the highest-energy step number, 31, corresponding to 220 keV/e.
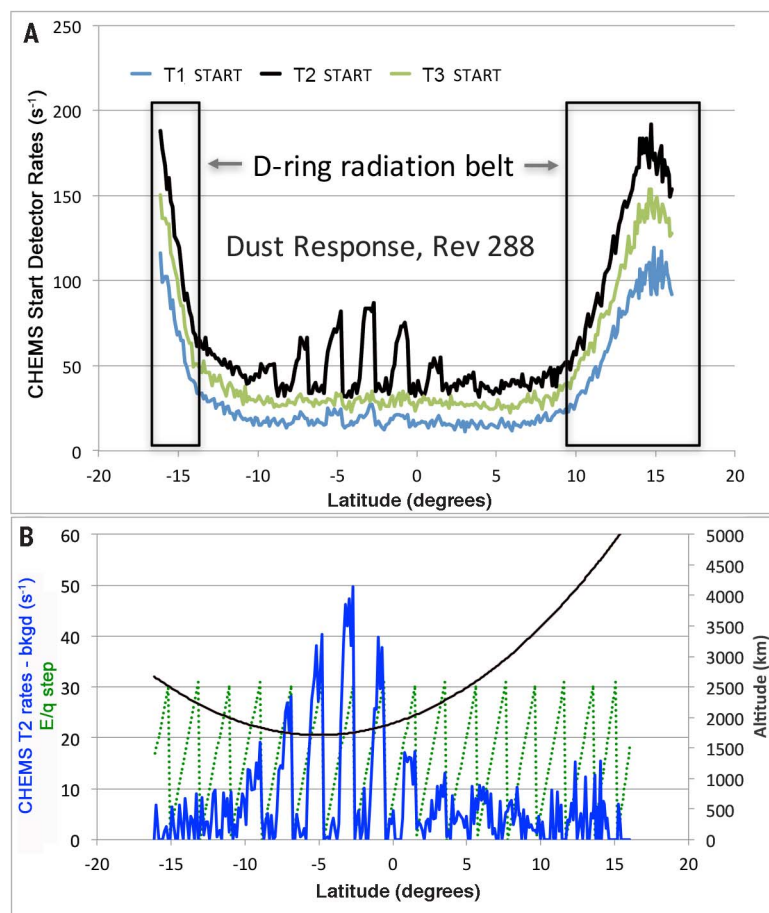
vector entered the CHEMS sensor aperture. The response of CHEMS to dust grains is similar to that of INCA (16) after the grains' transit of CHEMS' electrostatic analyzer (ESA). Dust particles with the requisite E/q to transit CHEMS' ESA reach the start foil, vaporize, and, if they are of sufficient energy, produce a plasma cloud at the foil exit plane that triggers the start MCP only. Electrons from this plasma trigger the start MCP but not the stop MCP. Measurements during the proximal orbits show that only starts, but no discernable

stops, are generated as the dust particle is destroyed by the impact with the start foil.

During the close approaches when dust was detected (revs 288 through 292, as well as rev 293, the final plunge by Cassini into Saturn's atmosphere), CHEMS telescope 2 (T2) was pointed in the corotation ram direction, defined as the spacecraft velocity added to the rotation velocity of the planetary atmosphere. CHEMS telescopes 1 (T1) and 3 (T3) were centered at ~60° above and below the ram direction. This spacecraft attitude

was designed to provide optimum pointing for the INMS instrument (25) but also proved beneficial for CHEMS because its T2 aperture was co-aligned with the INMS FOV. Although the CHEMS FOV faced the ram vector during the rolls in revs 273 and 284, that geometry was only reached well past the equator where dust densities were lower and the radiation backgrounds higher at the higher altitude of those periapses.

The start rates for the ram-facing T2 telescope produced a well-defined signal (Fig. 6). The weak

signal seen in T1 is not well understood (*16*). There was no discernable signal in any of the CHEMS stop rates or solid-state detector rates, other than the backgrounds from penetrating particles. This region is devoid of any trace of energetic charged particles in the design range for CHEMS or INCA (*18*), and at this altitude, the intensity of the energetic radiation belt particles is also very low, resulting in background low enough to measure this relatively weak signal. We have inferred that the response in T2 was produced by positively charged grains with $E/q > 40$ keV/e, up to the maximum measurable value of 220 keV/e. Similar responses were recorded by CHEMS during transits of the icy dust plumes emanating from the south polar region of Enceladus and in measurements fitting the Cassini CAPS instrument (*13*). See (*16*) for additional discussion of the CHEMS response to dust.

This energy-per-charge threshold, combined with the assumption that the grains are singly charged and that their energy is dominated by the ram velocity, indicates that the grains have masses of 8000 to 40,000 u. Assuming the density of water ice (this does not assume the grains are ice, as many organics have similar densities) implies radii of ~1.2 to 2.3 nm.

Figure 7 shows the CHEMS data for the six revs for which the pointing allowed CHEMS to face the dust ram vector in T2 and at low enough altitude so that the dust signal was above the radiation backgrounds in the CHEMS start rates. As before, we fitted the data with an exponential function of altitude, which results in a scale height of approximately 180±30 km. For revs 288, 289, and 290, an additional Gaussian in latitude is applied as a multiplying function to achieve an adequate description of the CHEMS dust profiles,

$$F(z, \lambda) = F_0 \exp[-(z - z_0)/H_D] \times \exp\{½[(\lambda - \lambda_0)/\sigma_\lambda]^2\} \quad (4)$$

where z is altitude above the 1-bar region, H_D is a scale height, and λ and σ_λ have the same meaning as in Eq. 1. The scale height H_D is held constant at 180 km, whereas the σ_λ parameter and amplitude F_0 were allowed to vary in this latitude-dependence fitting process. The z_0 was fixed at 1700 km, and λ_0 was set to 0.17° on the basis of the INCA measurements but is consistent with zero. Over several orbits, the σ_λ parameter changed from ~5° to 20° (Fig. 7). The profiles evolved from requiring an equatorial bias to being independent of latitude. The peak amplitudes of the profiles varied by up to a factor of two from one rev to the next, a much larger variability than that observed near 3000 km by INCA. The uncertainties in these parameters are similar to those indicated for Eqs. 1 and 2.

CHEMS measured positive dust particles near 1700-km altitude, which we compared to the dust measured by INCA near 3000 km. INCA made measurements for three revs, and these measurements were highly repeatable, with variations in density <5% and identical spatial and velocity distributions. Although sub-Keplerian, the velocities measured by INCA were well above the atmospheric (corotation) velocity and confined to a region close to the equator. CHEMS has a smaller 4° acceptance angle in the direction discriminating between ram velocities consistent with corotation or Keplerian and so would have detected nothing if the dust velocity differed from the corotation speed (for which the spacecraft attitude was tuned) by more than about 10%. This, and the observation that altitude dependence is well characterized by an exponential with a scale height (~180 km) similar to that of the atmosphere itself, leads to the conclusion that at this altitude, the transport is diffusive in nature, at least for positive dust grains.

We infer that the dust seen by INCA at high altitude is the source for the dust seen by CHEMS at lower altitude (with the original source being the D-ring, and likely the D68 ringlet). Why does the distribution vary in apparent density at the lower altitude? Continuity of mass flux suggests that it should be fairly constant, at least when integrated over all latitudes. Clearly the transport changes appreciably between 3000 and 1700 km, and asymmetries in the neutral atmosphere density relative to the equator may be involved. With more solar input north of the equator
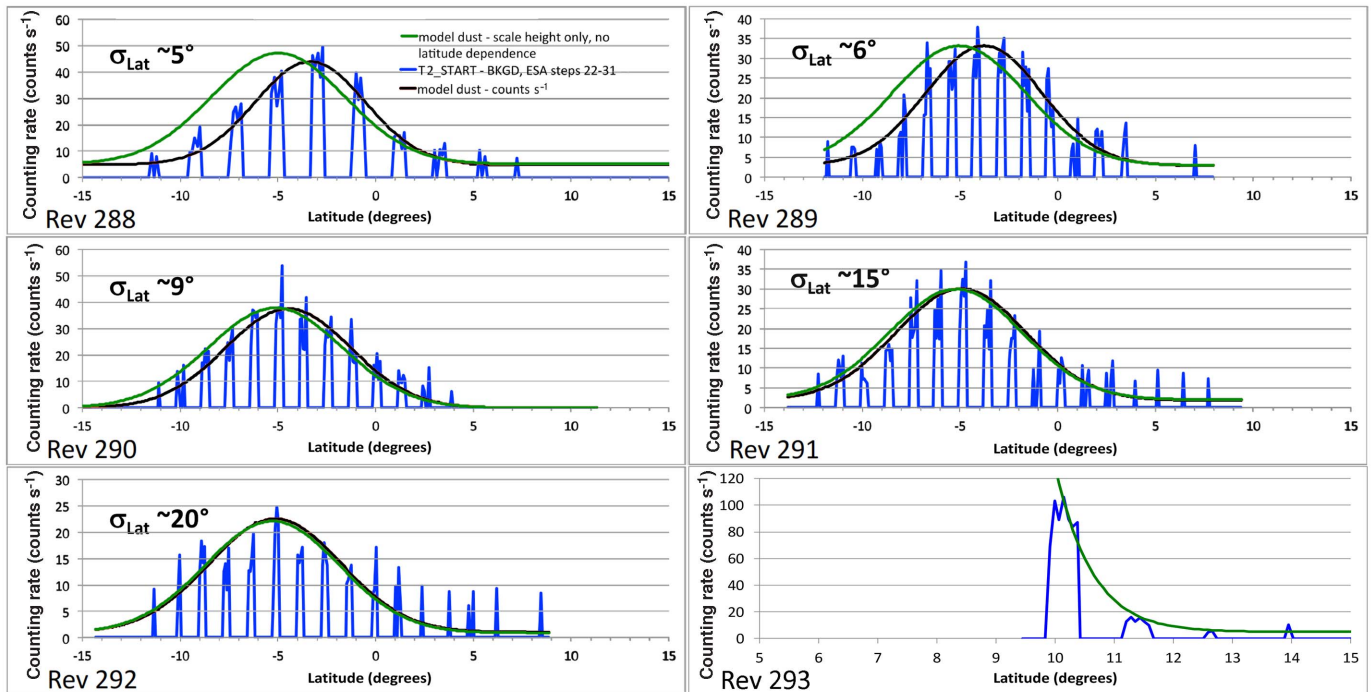
Fig. 7. CHEMS dust counting rates for six revs. Charged dust counting rates for all six revs for which CHEMS measured the charged dust signal. These include rev 293, the final plunge into Saturn's atmosphere, which reached higher rates at the lowest altitude and stopped when communication with the spacecraft was lost. Only the data at and above ESA step 19 (~40 keV/e, or about 8000 u/e) are shown, to reduce confusion introduced by noise or low-efficiency measurements in the lower ESA steps. Two model fits to the data are shown for each rev except rev 293. In green are exponentials in altitude with a common scale height of 180 km and varying amplitudes to match the measured rates. In black are the same exponentials but multiplied by a Gaussian in latitude centered at the equator with width σ_{Lat} as indicated (with amplitude again adjusted to fit the data). Revs 291 and 292 are consistent with, but do not strongly constrain, the equator-centered Gaussian. Rev 293, with the steep entry into the atmosphere, did not constrain equator weighting. The first three revs favor the Gaussian fit. A small constant level was added to each fit to accommodate the wings (ingress and egress).

during this part of Saturn's year, the atmospheric scale heights may be higher in the north than in the south. Winds in the thermosphere may further alter the latitude density profiles of H and H_2. Although there is strong evidence for the low-altitude dust being biased toward the equator on revs 288 and 289, that signature is less evident or even absent in later revs.

Our kinetic, collisional model of the dust transport qualitatively matches the observed densities, latitude distribution, and velocities but does not quantitatively match. We have so far treated the dust as neutral in the transport calculations; however, this is not an adequate assumption for the charged component of the dust. As discussed above, for the relevant size range (~1- to 3-nm radius), the most probable charge state at higher altitudes is neutral and the charging times are hours to days. Even deeper in the ionosphere, at the 1700-km level, this is still the case, although the charging times (especially the times to charge negatively) decrease to less than an hour.

Comparing dust at different altitudes

At higher altitudes (3000 km and above), any charged dust particles are picked up by the corotational electric field felt by the particles owing to the difference between their Keplerian velocity and the corotation velocity, which forms the rest frame of the magnetic field. The combined gravitational and electromagnetic forces quickly destabilize the dust particles orbits, so they precipitate along the magnetic field into the atmosphere, away from the equator (30–32). The INCA measurements appear to measure this

population, as the higher-latitude wings of the equatorially confined population of dust is nearly 100% charged, that is, very likely the charged particles that have recently converted from neutrals and are in the process of precipitating along the magnetic field.

At the lower altitudes, collisions with the neutral atmosphere will damp electromagnetic effects, and charged dust is not expected to move guided by the field for any great distance. However, the dust-grain charge states may still play a role in their transport. Our transport calculations predict reduced, but still substantial, vertical velocities for the neutral dust. Having been scattered to the point that their net horizontal velocity is zero in the atmosphere rest frame, they still are influenced by the force of gravity. At these lower altitudes of ~1700 km, the atmospheric density is sufficiently high that the dust particles no longer free fall but should reach a terminal velocity. That terminal velocity is still on the order of 1 to 2 km s^{-1} in the altitude range from 1800 to 1700 km where diffusive spatial distributions ordered by scale height are observed. Using the dust flux continuity calculations (Fig. 6), we predict a dust density of ~1 or 2 cm^{-3} at ~1700 km, whereas INMS measurements (25) suggest dust densities of several thousand per cubic centimeter. We cannot address this directly with the CHEMS data because this instrument only measures the positively charged fraction of the dust.

Charging can also slow the dust as it falls through the atmosphere. Charged dust not only experiences the drag force of kinetic collisions with atmospheric atoms and molecules but will

also be stopped by electromagnetic $V \times B$ forces, where V is particle velocity relative to magnetic field B. Previous studies of these forces on charged grains (33) were restricted to much larger grains, for which the ratio of electromagnetic to gravitational forces is very different. If sufficient dust grains are charged for sufficient lengths of time, much higher densities could build up. The charging times shorten as the ionospheric density increases, so a larger dust fraction may charge at low altitudes relative to the higher altitudes. Most of that charging will be negative, and CHEMS cannot measure such grains.

Comparison with other instruments

The MIMI-CHEMS measurements can be readily compared to those from the RPWS and INMS instruments, which made observations concurrent with those of MIMI during the closest approaches (25). RPWS obtained electron density, and INMS directly measured light ion densities of H$^+$ and H$_3^+$ during closest approach.

Figure 8A shows the relatively close agreement between the light ion density and the plasma electron density farther from the equator (which also corresponds to higher altitudes, see Fig. 6) and the strong departure of the two densities near the equator (lower altitudes). In Fig. 8B, we take the difference between the electron density and the light ion density and overplot the CHEMS T2 dust counting rate for the E/q steps with high dust efficiencies, multiplied by an arbitrary constant. Charge conservation requires that the total ion density equal the total electron density N_e. However, near closest approach, N_e
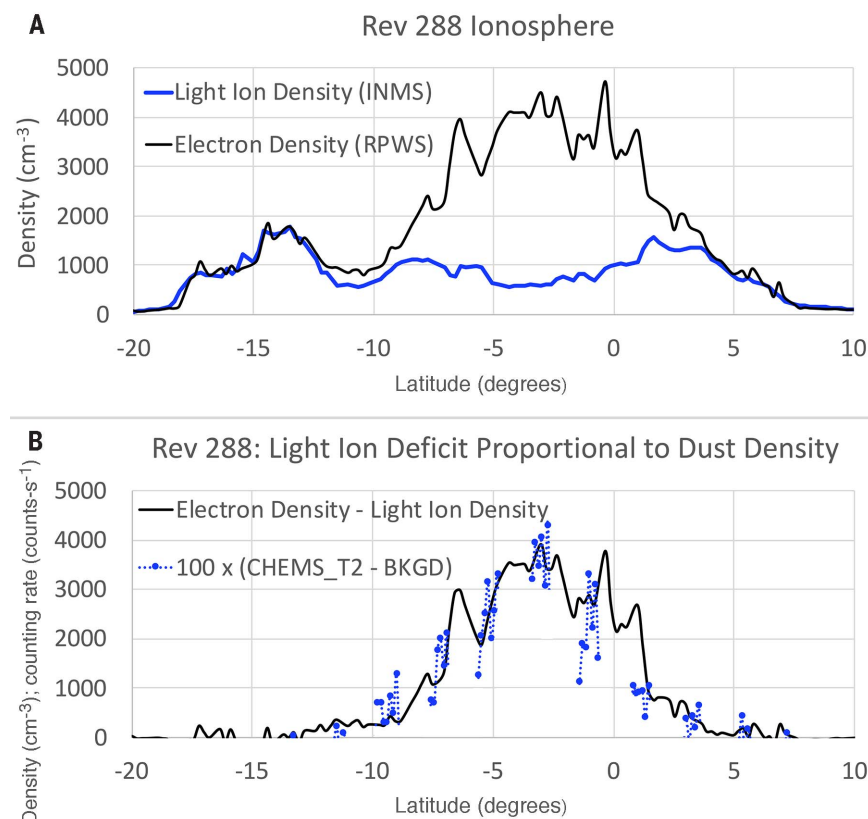
Fig. 8. Rev 288 MIMI-CHEMS, RPWS, and INMS results compared. (**A**) Electron density (black curve) as determined by the RPWS instrument by using characteristic plasma resonances and light ion (H$^+$ and H$_3^+$) density as measured by INMS. The difference between the two curves indicates the presence of a heavier ion constituent not measured by INMS under these conditions, which must be present to achieve overall charge neutrality. (**B**) The difference in the two densities shown in (A), with the CHEMS T2 rates (background corrected), multiplied by a constant and overplotted in blue dots. The correspondence between the two curves indicates a proportionality between the dust measured by CHEMMS and the plasma heavy ion population.

lies above the total light ion density of H^+ and H_3^+ by nearly an order of magnitude. The missing ions are thought to be more massive molecules that are too energetic to be detected by INMS at this ram velocity (*25*). It is evident that the density of the positive dust grains measured by CHEMS is directly proportional to the heavier ion densities required for charge neutrality. The CHEMS measurement does not address this missing positive charge; the CHEMS counting rates—when converted to a positive dust grain density accounting for the CHEMS geometric factor; ESA passband, which is an estimate of efficiency; and the ram velocity—yield a dust-grain density of about 0.002 cm^{-3}, which is far too low to account for the required positive charge density (>4000 cm^{-3} at the peak). But the observation that the positive dust-grain density scales linearly with the undetected ionospheric positive charge suggests a common transport path and a common source in the D-ring [compare (*25*)].

Summary and conclusions

Dust grains with masses from ~8000 to >40,000 u have been detected in situ inside Saturn's rings. Modeling their distribution indicates that dust is precipitating into Saturn's equatorial atmosphere directly from the inner edge of the D-ring and probably from the D68 ringlet. This dust population is tightly confined near the equator at higher altitudes (~3000 km), as directly observed by the MIMI-INCA instrument. At this ~3000-km altitude, the dust velocity distribution is measured to be sub-Keplerian speed (between circular Keplerian and about 70% Keplerian, peaking at 85% Keplerian). The dust population has a charged component that comprises ≥20% of the population in the equatorial plane, reaching 100% of the population on the higher-latitude (beyond ±~5°) wings of the spatial distribution, owing to spatial filtering of the original population by the Lorentz force.

The dust is precipitating through atmospheric drag, a process previously invoked for the interaction between Uranus's ring and its atmosphere (*34*). At the highest altitudes, the nature of this drag is kinetic collisions between the dust grains and exospheric H. Many of these collisions (hundreds per dust particle) will result in H atoms with escape or near-escape velocities, in a specific direction (tangential to the atmosphere at the point of collision and close to the equator in direction) and at altitudes where the mean free path is very large. This is expected to produce a durable, high-altitude, equatorially confined escaping population of neutral H, as well as some H_2 by the same process. An analogy of this population is sawdust ejected from a circular saw as it cuts through wood. We suggest that this scenario is consistent with the previously observed equatorial hydrogen plume (*35*).

MIMI-CHEMS measured the positively charged fraction of the dust at lower altitudes (~1700 km), where the densities are presumably much higher. The dust at this altitude is organized by a scale height, suggesting diffusion-dominated transport, unlike the dust observations at 3000 km. The source for this dust still appears to be near the

equator. Kinetic transport calculations predict lower densities than are inferred for the dust population at this altitude (*25*). We argue that dust charging plays a role in slowing the fall of the dust through the atmosphere. The CHEMS densities of positively charged dust grains are shown to be directly proportional to the heavy positive-ion population that is inferred from the difference between the light ion density and the electron density measured at the lowest altitudes on revs 288 and 292.

REFERENCES AND NOTES

1. J. E. P. Connerney, J. H. Waite, New model of Saturn's ionosphere with an influx of water from the rings. *Nature* **312**, 136–138 (1984). doi: 10.1038/312136a0
2. J. E. P. Connerney, Magnetic connection for Saturn's rings and atmosphere. *Geophys. Res. Lett.* **13**, 773–776 (1986). doi: 10.1029/GL013i008p00773
3. T. G. Northrop, J. R. Hill, Stability of negatively charged dust grains in Saturn's ring plane. *J. Geophys. Res.* **87**, 6045–6051 (1982). doi: 10.1029/JA087iA08p06045
4. T. G. Northrop, J. R. Hill, The inner edge of Saturn's B ring. *J. Geophys. Res.* **88**, 6102–6108 (1983). doi: 10.1029/JA088iA08p06102
5. J. Connerney, Saturn's ring rain. *Nature* **496**, 178–179 (2013). doi: 10.1038/496178a; pmid: 23579675
6. J. O'Donoghue et al., Redetection of the ionospheric H_3^+ signature of Saturn's "ring rain". *Geophys. Res. Lett.* **44**, 11,762–11,739 (2017). doi: 10.1002/2017GL075932
7. A. J. Kliore et al., Midlatitude and high-latitude electron density profiles in the ionosphere of Saturn obtained by Cassini radio occultation observations. *J. Geophys. Res.* **114**, A04315 (2009). doi: 10.1029/2008JA013900
8. A. J. Kliore et al., The ionosphere of Saturn as observed by the Cassini Radio Science System. *Geophys. Res. Lett.* **41**, 5778–5782 (2014). doi: 10.1002/2014GL060512
9. L. Moore, N. R. Iverson, D. Cohen, Conditions for thrust faulting in a glacier. *J. Geophys. Res.* **115**, A11317 (2010). doi: 10.1029/2009JF001307
10. J.-E. Wahlund et al., In situ measurements of Saturn's ionosphere show that it is dynamic and interacts with the rings. *Science* **359**, 66–68 (2018). doi: 10.1126/science.aao4134; pmid: 29229651
11. Y. Dong, T. W. Hill, S.-Y. Ye, Characteristics of ice grains in the Enceladus plume from Cassini observations. *J. Geophys. Res.* **120**, 915–937 (2015). doi: 10.1002/2014JA020288
12. N. Krupp et al., Energetic electron measurements near Enceladus by Cassini during 2005–2015. *Icarus* **306**, 256–274 (2018). doi: 10.1016/j.icarus.2017.10.022
13. W. T. Hill et al., Charged nanograins in the Enceladus plume. *J. Geophys. Res.* **117**, A05209 (2012). doi: 10.1029/2011JA017218
14. A. J. Coates et al., Discovery of heavy negative ions in Titan's ionosphere. *Geophys. Res. Lett.* **34**, L22103 (2007). doi: 10.1029/2007GL030978
15. S. M. Krimigis et al., Magnetosphere Imaging Instrument (MIMI) on the Cassini Mission to Saturn/Titan. *Space Sci. Rev.* **114**, 233–329 (2004). doi: 10.1007/s11214-004-1410-8
16. Instrument descriptions and data analyses methods are available as supplementary materials.
17. M. K. Dougherty et al., The Cassini Magnetic Field Investigation. *Space Sci. Rev.* **114**, 331–383 (2004). doi: 10.1007/s11214-004-1432-2
18. E. Roussos et al., A radiation belt of energetic protons located between Saturn and its rings. *Science* **362**, eaat1962 (2018).
19. C. K. Goertz, Dusty plasmas in the solar system. *Rev. Geophys.* **27**, 271–289 (1989). doi: 10.1029/RG027i002p00271
20. O. Havnes, G. E. Morfill, F. Melandsø, Effects of electromagnetic and plasma drag forces on the orbit evolution of dust in planetary magnetospheres. *Icarus* **98**, 141–150 (1992). doi: 10.1016/0019-1035(92)90215-S
21. L. J. Gelinas et al., Mesospheric charged dust layer: Implications for neutral chemistry. *J. Geophys. Res.* **110**, A01310 (2005). doi: 10.1029/2004JA010503
22. C. K. Goertz, G. E. Morfill, A new instability of Saturn's ring. *Icarus* **74**, 325–330 (1988). doi: 10.1016/0019-1035(88)90044-9
23. J. H. Waite Jr. et al., The Cassini Ion and Neutral Mass Spectrometer (INMS) Investigation. *Space Sci. Rev.* **114**, 113–231 (2004). doi: 10.1007/s11214-004-1408-2
24. T. T. Koskinen et al., Saturn's variable thermosphere from Cassini/UVIS occultations. *Icarus* **260**, 174–189 (2015). doi: 10.1016/j.icarus.2015.07.008

25. J. H. Waite Jr. et al., Chemical interactions between Saturn's atmosphere and its rings. *Science* **362**, eaat2382 (2018).
26. T. E. Cravens et al., The ion composition of Saturn's equatorial ionosphere as observed by Cassini. *Geophys. Res. Lett.* 10.1029/2018GL077868 (2018).
27. A. R. Poppe, An improved model for interplanetary dust fluxes in the outer Solar System. *Icarus* **264**, 369–386 (2016). doi: 10.1016/j.icarus.2015.10.001
28. H.-W. Hsu et al., In situ collection of dust grains falling from Saturn's rings into its atmosphere. *Science* **362**, eaat3185 (2018).
29. M. M. Hedman, M. R. Showalter, A new pattern in Saturn's D ring created in late 2011. *Icarus* **279**, 155–165 (2016). doi: 10.1016/j.icarus.2015.09.017
30. H.-W. Hsu et al., Cassini dust stream particle measurement during the first three orbits at Saturn. *J. Geophys. Res.* **116**, A08213 (2011). doi: 10.1029/2010JA015959
31. S. Kempf et al., High-velocity streams of dust originating from Saturn. *Nature* **433**, 289–291 (2005). doi: 10.1038/nature03218; pmid: 15662418
32. W.-H. Ip, An update on the ring exosphere and plasma disc of Saturn. *Geophys. Res. Lett.* **32**, L13204 (2005). doi: 10.1029/2004GL022217
33. W.-H. Ip, C.-M. Liu, K.-C. Pan, Transport and electrodynamical coupling of nano-grains ejected from the Saturnian rings and their possible ionospheric signatures. *Icarus* **276**, 163–169 (2016). doi: 10.1016/j.icarus.2016.04.004
34. A. L. Broadfoot et al., Ultraviolet spectrometer observations of Uranus. *Science* **233**, 74–79 (1986). doi: 10.1126/science.233.4759.74; pmid: 17812892
35. D. E. Shemansky, X. Liu, H. Melin, The Saturn hydrogen plume. *Planet. Space Sci.* **57**, 1659–1670 (2009). doi: 10.1016/j.pss.2009.05.002
36. D. G. Mitchell, Dust trajectory model. Zenodo (2018); doi: 10.5281/zenodo.1255268.

RESEARCH ARTICLE SUMMARY

GAS GIANT PLANETS

Chemical interactions between Saturn's atmosphere and its rings

J. H. Waite Jr.*†, R. S. Perryman†, M. E. Perry†, K. E. Miller†, J. Bell, T. E. Cravens, C. R. Glein, J. Grimes, M. Hedman, J. Cuzzi, T. Brockwell, B. Teolis, L. Moore, D. G. Mitchell, A. Persoon, W. S. Kurth, J.-E. Wahlund, M. Morooka, L. Z. Hadid, S. Chocron, J. Walker, A. Nagy, R. Yelle, S. Ledvina, R. Johnson, W. Tseng, O. J. Tucker, W.-H. Ip

INTRODUCTION: Past remote observations of Saturn by Pioneer 11, Voyager 1 and 2, Earth-based observatories, and the Cassini prime and solstice missions suggested an inflow of water from the rings to the atmosphere. This would modify the chemistry of Saturn's upper atmosphere and ionosphere. In situ observations during the Cassini Grand Finale provided an opportunity to study this chemical interaction.

RATIONALE: The Cassini Grand Finale consisted of 22 orbital revolutions (revs), with the closest approach to Saturn between the inner D ring and the equatorial atmosphere. The Cassini Ion Neutral Mass Spectrometer (INMS) measured the composition of Saturn's upper atmosphere and its chemical interactions with material originating in the rings.

RESULTS: Molecular hydrogen was the most abundant constituent at all altitudes sampled. Analysis of the atmospheric structure of H_2 indicates a scale height with a temperature of 340 ± 20 K below 4000 km, at the altitudes and near-equatorial latitudes sampled by INMS.

Water infall from the rings was observed, along with substantial amounts of methane, ammonia, molecular nitrogen, carbon monoxide, carbon dioxide, and impact fragments of organic nanoparticles. The infalling mass flux was calculated to be between 4800 and 45,000 kg s^{-1} in a latitude band of 8° near the equator.

The interpretation of this spectrum is complicated by the Cassini spacecraft's high velocity of 31 km s^{-1} relative to Saturn's atmosphere. At this speed, molecules and particles have 5 eV per nucleon of energy and could have fragmented upon impact within the INMS antechamber of the closed ion source. As a result, the many organic compounds detected by INMS are very likely fragments of larger nanoparticles.

Evidence from INMS indicates the presence of molecular volatiles and organic fragments in the infalling material. Methane, carbon monoxide, and nitrogen make up the volatile inflow, whereas ammonia, water, carbon dioxide, and organic compound fragments are attributed to fragmentation inside the instrument's antechamber of icy, organic-rich grains. The observations also show evidence for orbit-to-orbit variations in the mixing ratios of infalling material; this suggests that the source region of the material is temporally and/or longitudinally variable, possibly corresponding to localized source regions in the D ring.

CONCLUSION: The large mass of infalling material has implications for ring evolution, likely requiring transfer of material from the C ring to the D ring in a repeatable manner. The infalling material can affect the atmospheric chemistry and the carbon content of Saturn's ionosphere and atmosphere. ■

Atmospheric composition
4.7×10^{-15} g cm^{-3}
- He
- Hydrogen

Ring composition
1.6×10^{-16} g cm^{-3}
Mass percent of ring material + 1σ

- CH_4 (16 ±3)%
- CO_2 (0.5±0.1)%
- CO/N_2 (20±3)%
- H_2O (24±5)%
- NH_3 (2.4±0.5)%
- Organics (37±5)%

Rev 290 Mass Influx: 6500-25,000 kg s^{-1}

Rev 290 Integrated counts per IP

Mass per charge (u)
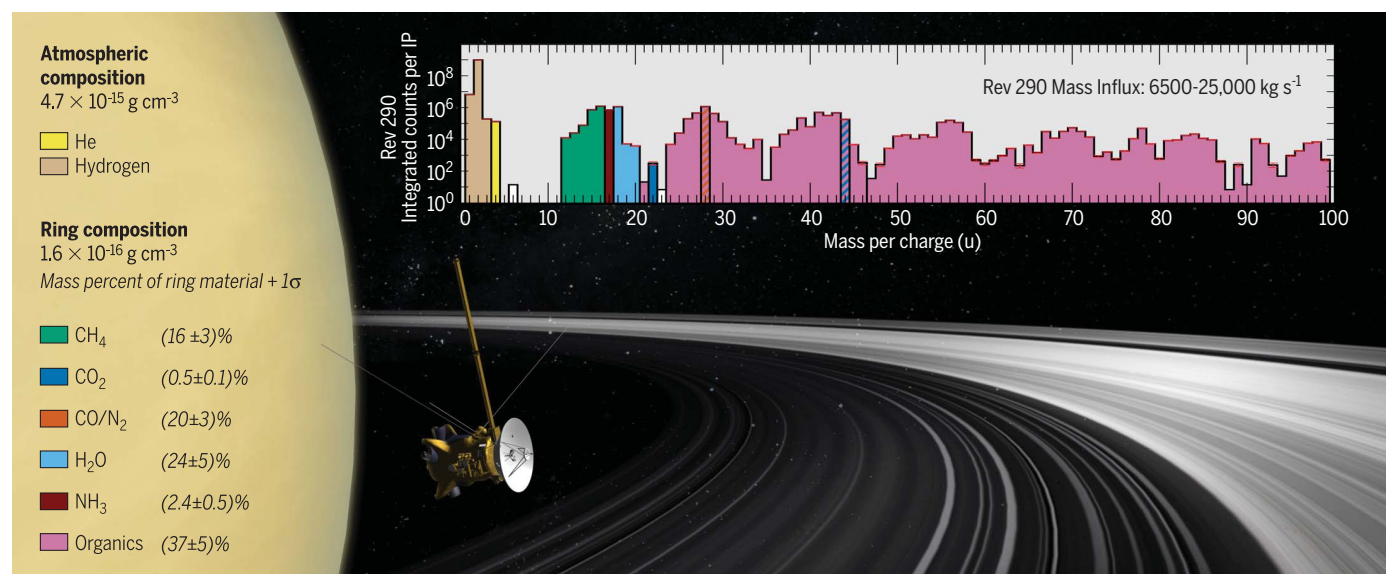
INMS mass spectra from the Grand Finale. The graphic depicts the Cassini spacecraft as it passes from north to south between Saturn and its rings. The inset spectrum shows the mass deconvolution of compounds measured by INMS on rev 290. The *x* axis is in units of mass per charge (u) and extends over the full mass range of INMS (1 to 99 u). The *y* axis is in counts per measurement cycle integrated over the closest-approach data. The mass influx rate for rev 290, derived from mass deconvolution of the rev-integrated spectrum, is shown as embedded text in the spectrum. The side panel gives the average of the mass deconvolution of revs 290, 291, and 292 in mass density units (g cm^{-3}). The composition of the ring-derived compounds in terms of percentage mass density is also shown.

RESEARCH ARTICLE

GAS GIANT PLANETS

Chemical interactions between Saturn's atmosphere and its rings

J. H. Waite Jr.[1]*†, R. S. Perryman[1]†, M. E. Perry[2]†, K. E. Miller[1]†, J. Bell[3,4],
T. E. Cravens[5], C. R. Glein[1], J. Grimes[1], M. Hedman[6], J. Cuzzi[7], T. Brockwell[1], B. Teolis[1],
L. Moore[8], D. G. Mitchell[2], A. Persoon[9], W. S. Kurth[9], J.-E. Wahlund[10], M. Morooka[10],
L. Z. Hadid[10], S. Chocron[1], J. Walker[1], A. Nagy[11], R. Yelle[12], S. Ledvina[13], R. Johnson[14],
W. Tseng[15], O. J. Tucker[4], W.-H. Ip[16]

The Pioneer and Voyager spacecraft made close-up measurements of Saturn's ionosphere and upper atmosphere in the 1970s and 1980s that suggested a chemical interaction between the rings and atmosphere. Exploring this interaction provides information on ring composition and the influence on Saturn's atmosphere from infalling material. The Cassini Ion Neutral Mass Spectrometer sampled in situ the region between the D ring and Saturn during the spacecraft's Grand Finale phase. We used these measurements to characterize the atmospheric structure and material influx from the rings. The atmospheric He/H_2 ratio is 10 to 16%. Volatile compounds from the rings (methane; carbon monoxide and/or molecular nitrogen), as well as larger organic-bearing grains, are flowing inward at a rate of 4800 to 45,000 kilograms per second.

Early modeling of Saturn's atmosphere/ionosphere coupling (*1*) prior to the first radio occultation measurements by Pioneer 11 adopted compositional constraints from Earth-based observations of the well-mixed saturnian atmosphere and used a range of values of the turbulent mixing and heating of the atmosphere based on past planetary observations. Depending on the chosen thermal profile and strength of turbulent mixing, a range of atmospheric and ionospheric conditions were deemed possible (*1*). The "nominal" model of the ionosphere predicted that protons were the primary ion and that the proton's slow radiative recombination reaction was the primary chemical loss pathway, indicating a density of the ionosphere reaching 10^5 cm^{-3}. The modeling suggested that ionospheric composition and structure could serve as a diagnostic of the composition, thermal structure, and turbulent mixing of the upper atmosphere. However, the first radio occultation measurements of the ionosphere made by Pioneer 11 on 1 September 1979 found a peak electron density an order of magnitude lower than predicted for moderate eddy mixing and a warm thermosphere (*2*). The discrepancy was not well understood at the time.

Voyager radio occultation measurements (*3*) supported the low peak ionospheric density and provided independent evidence that the low peak ionospheric density extended into the night side (*4*). It was suggested (*5*, *6*) that water from the rings flowing into the atmosphere could chemically convert protons to molecular H_3O^+ ions at a rate that reproduced the peak electron densities observed by Pioneer and Voyager. The globally averaged water influx rate was estimated to be 4×10^7 cm^{-2} s^{-1}, with localized influx as high as 2×10^9 cm^{-2} s^{-1} (*6*). This proposal of water influx was bolstered by further modeling (*7*, *8*) and ground-based observations of H_3^+ in the Saturn ionosphere (*9*).

The 59 Cassini radio occultations carried out over the course of its prime and solstice missions provided additional data on the ionospheric structure (*8*, *10*). Low latitudes—within 20° of the equator—have lower ionospheric density. Water influx from the rings is a strong candidate to explain the observed latitudinal variations; modeling (*8*) indicates that water influx with a Gaussian distribution about the equator and a peak flux of 5×10^6 cm^{-2} s^{-1} can match the observations obtained prior to the Grand Finale time period.

Observations during the Cassini Grand Finale

The Grand Finale phase of the Cassini mission began on 22 April 2017 with a final close flyby of Titan that diverted the spacecraft to fly 22 times between the planet and its rings. The mission ended on 15 September 2017 after another gravitational deflection by Titan that sent the spacecraft plunging into Saturn's dense atmosphere. Cassini's trajectory between the innermost D ring and the planet allowed in situ coverage of the equatorial ring-atmosphere interaction. Flybys varied in altitude between 1360 and 4000 km above the atmosphere's 1-bar pressure level in three groups (Fig. 1).

The Grand Finale objectives included measuring Saturn's atmospheric H_2 and He and searching for compounds such as water that might indicate an interaction of the upper atmosphere with the main rings. The Ion Neutral Mass Spectrometer (INMS) (*11*) measurements were made several hundred kilometers above the homopause, the level at which turbulent or eddy mixing and molecular diffusion are equal (Fig. 1). In this region, molecular diffusion produces a mass-dependent separation: Lighter compounds have a larger vertical extent due to Saturn's gravity. Prior to the Grand Finale, models and occultation data (*12*, *13*) suggested that H_2, He, and HD were the only neutrals in the upper atmosphere that would be measurable by INMS. The heavier molecule methane was predicted to be present at the ~0.1% level in the well-mixed lower atmosphere (below 1000 km; Fig. 1), but below the detection limits of INMS at the high altitudes sampled by Cassini (*12*, *13*).

Atmospheric composition

INMS obtained measurements of neutral molecules using the Closed Source Neutral (CSN) mode (*11*). INMS operated in a survey mode, acquiring mass spectral data at every mass within its range [1 to 99 atomic mass units (u) at resolution of 1 u], with a repetition rate of 9.5 s (~300 km along track) or 4.5 s (~150 km) when ionospheric data were or were not being acquired sequentially, respectively.

Figure 1 indicates the three altitude regions between the atmosphere and the rings where measurements were made by INMS. Figure 2 shows data for the major atmospheric components H_2, ^{4}He, and HD + ^{3}He, separated into those three altitude bands. HD and ^{3}He have the same atomic mass, so they cannot be measured separately. Because of the geometry of the orbit (rev) for each altitude region, there is a strong correlation between altitude and latitude. The closest approach to Saturn occurred at 4° to 6°S and not at the equator. This allowed us to determine whether the observed compounds are better correlated with a neutral flux from the rings or with Saturn's atmospheric structure (see below). All orbits occurred near local solar noon. This geometry provides complementary

information to the Pioneer, Voyager, and Cassini radio occultations, which probed near the dusk and dawn terminators. The Cassini orbits covered a range of Saturn rotational longitudes (Fig. 2). For the low-altitude observations (Fig. 2C), the differences in the shape of the H_2, HD + ^{3}He, and ^{4}He altitude profiles are due to changes in their respective abundances and indicate the effects of diffusive separation via gravity.

H_2 and He data were examined to determine the variability due to altitude, latitude, and orbit, the latter representing temporal and/or longitudinal effects. For each orbit, raw count data were converted to densities (14) and an average altitude and density were calculated every 5° latitude using the 11 data points closest in latitude (Fig. 3). Local scale heights were then fitted at each latitude from surrounding data having densities within a factor of $1/(2e)$ higher or lower than the average density, one e-fold about on the latitude of interest.

In addition to local scale heights, a common scale height was calculated using data across all observed altitudes, latitudes, and revs. This was facilitated because altitude and latitude are highly correlated and do not vary much from one orbit to the next. With the exception of 5°N latitude, all altitude/latitude bins are consistent with this common scale height (H), given by $H = kT/(mg)$, where k is Boltzmann's constant, T is the H_2 gas temperature, g is the local acceleration due to gravity, and m is the mass of the gas molecule, H_2 in this case. The common scale height represented in Fig. 3 corresponds to a temperature of 340 ± 20 K. The slopes of the local scale height at 5°N are steeper than the common scale height fit, suggesting a higher temperature. The scale height correspondence between the various altitudes/latitudes sampled, the times/longitudes sampled (on the different revs), and the locally determined scale heights calculated at each point show little evidence for major temporal or spatial atmospheric variation in the region sampled by Cassini during the Grand Finale (15).

The analysis of the INMS measurements indicates an atmosphere that has a more complex mass spectrum than predicted by models. There is evidence for both volatile compounds and fragments derived from nanograin impacts in the INMS antechamber. We conclude that the nanograins are sourced from the ring plane. Figure 4 shows the latitude distributions of mass 15 u and mass 28 u in the three distinct altitude ranges. Mass 15 u is the CH_3^+ fragment of methane (mass 16 u is not used because of interference from mass 16 oxygen from water electron impact dissociative ionization), whereas mass 28 u may be derived from N_2, CO, and/or C_2H_4. Mass deconvolution for revs 290, 291, and 292 indicates that less than 10% of the 15 u peak is derived from dissociative fragments from ammonia or heavy organics. From deconvolution of these same spectra, we attribute the fractional contributions of the 28 u signal to be 28% to 58% (average 45%) CO, 41% to 50% (average 46%) N_2, and 0% to 26% (average 9%) C_2H_4, an organic formed from fragmentation of nanograins.

Discrimination between volatile and nanograin-derived signals is achieved by comparing the data in Fig. 4 at the three separate altitude ranges. The uppermost altitude range shows that both mass 28 u and mass 15 u reach a maximum abundance at the equator rather than at the closest-approach latitude (~5°S) with a very large spike for mass 28 u at the equator, which we attribute to the C_2H_4 portion of the peak. The broad latitude extent of the distributions indicates a distributed volatile source in the ring plane. The mass 28 u signal is consistent with C_2H_4 derived from nanograins that have a very narrow Gaussian distribution peaked within 1° of the ring plane at this altitude (~3500 km), comparable to the nanograin distribution seen by the Magnetospheric Imaging Instrument (MIMI) investigation (16).
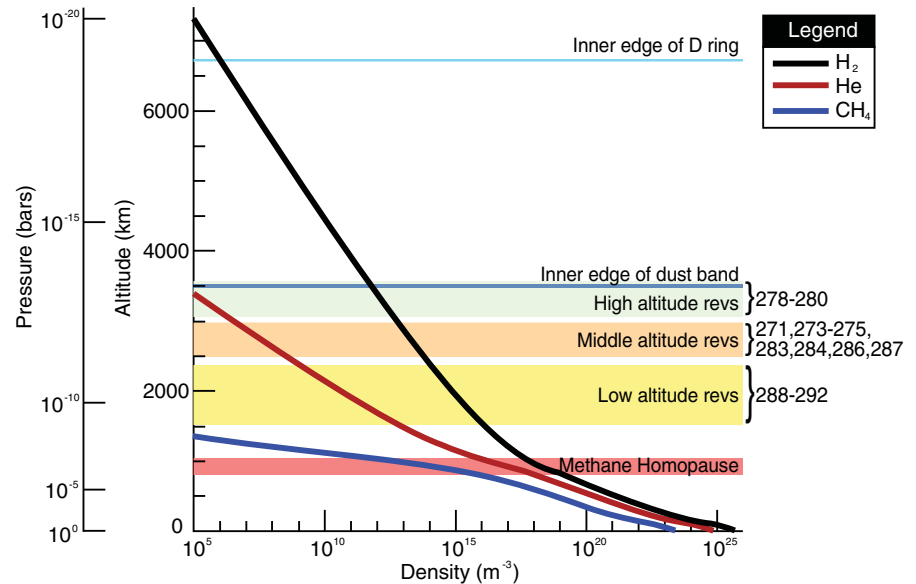
Fig. 1. Predicted structure of Saturn's atmosphere. Density profiles for the major atmospheric constituents H_2, He, and CH_4 (thick curves) were calculated using the hydrostatic model outlined in (29). This model was originally developed for Titan but has been modified for Saturn by altering the gravity calculation by including the centrifugal acceleration at the equator, assuming the mean thermal structure determined by Cassini observations (35), updating the molecular and eddy diffusion coefficients (7), and adjusting the composition at the lower boundary to 13.55% He, 0.45% CH_4, and 86% H_2 (35). Overlain are the altitudes at closest approach for the Grand Finale orbits (revs), which form three groups.

Table 1. Mass influx values and composition of inflowing ring material. Density and composition values are the average of revs 290, 291, and 292.

	Mass influx (kg s^{-1})	
Rev number	Minimum	Maximum
290	6,500	25,000
291	12,000	45,000
292	4,800	18,000

Source	Density	Source composition (mass fraction ± 1σ uncertainty)
Saturn's atmosphere	4.7×10^{-15} g cm^{-3}	Hydrogen: 99.95 ± 0.01%
		Helium: 0.05 ± 0.01%
Inflowing ring material	1.6×10^{-16} g cm^{-3}	Methane: 16 ± 3%
		Ammonia: 2.4 ± 0.5%
		Water: 24 ± 5%
		CO/N_2: 20 ± 3%
		Carbon dioxide: 0.5 ± 0.1%
		Organics: 37 ± 5%

Both the mass 15 u and mass 28 u compounds also have a component that has begun to spread in latitude toward the closest-approach latitude (4° to 6°S) even at the highest altitudes measured. This is more obvious for mass 15 u than for mass 28 u. This spreading indicates that collisions between the volatile ring compo-

nents and the saturnian atmosphere (H_2 and He) are beginning to "diffusively couple" the volatiles into the atmosphere as they flow from the rings. At the measured atmospheric densities for this altitude range, diffusive coupling can only occur with molecular volatiles and not heavier nanograins. The latter undergo a

smaller relative momentum exchange with the predominantly light hydrogen atmosphere. In the lower-altitude bands, the equatorial peaks disappear as both volatiles and nanograins are diffusively coupled into the atmosphere at the correspondingly higher atmospheric densities. This can be seen for the lower-altitude bands as
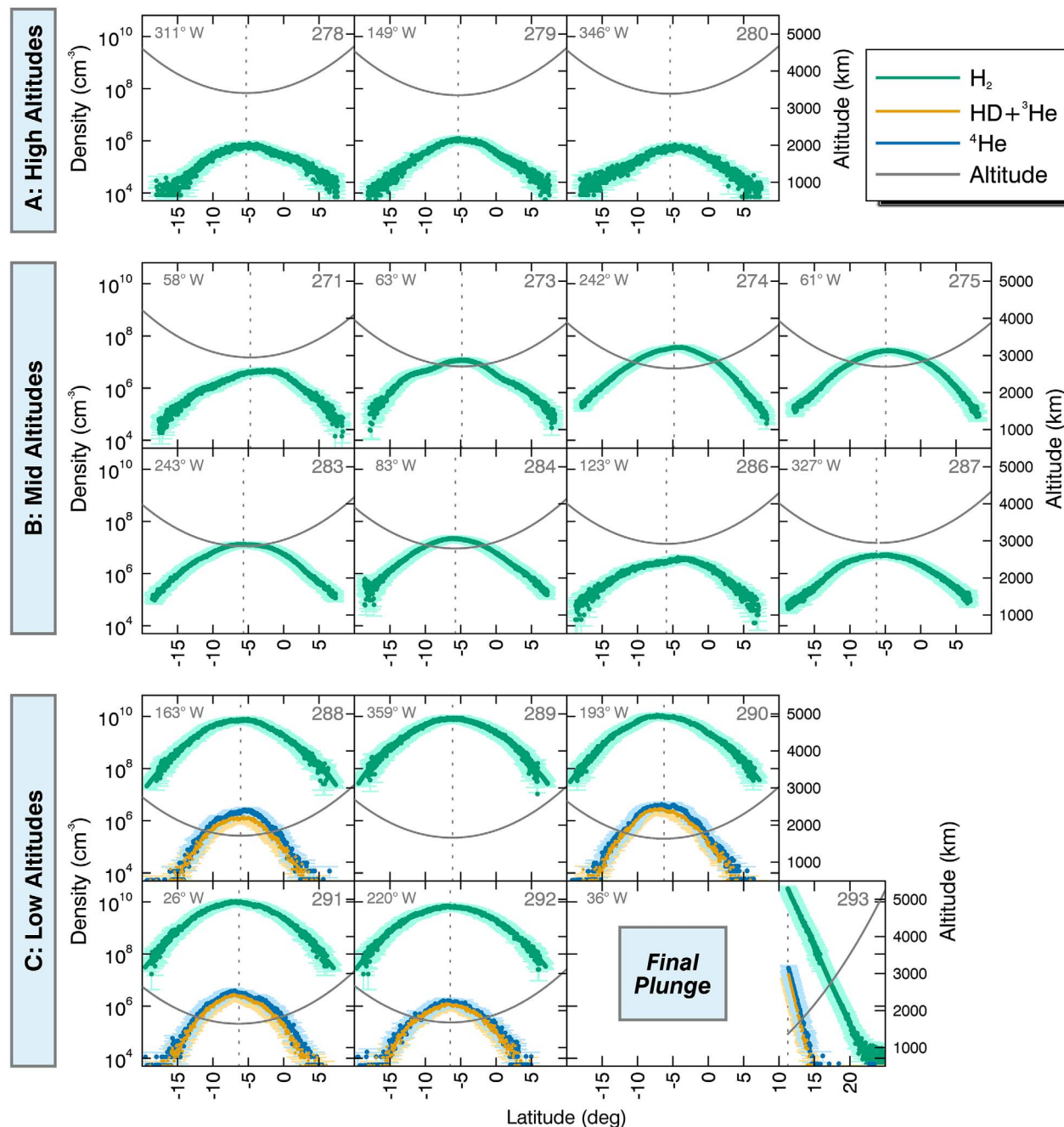
Fig. 2. Major atmospheric gases observed by INMS. The measured densities (points, with shading representing the 1σ uncertainty) of H_2, ^{4}He, and HD + ^{3}He are shown along the spacecraft trajectory for all INMS observations during the Grand Finale. (**A** to **C**) Data are grouped into the three altitude bands denoted in Fig. 1. The signal-to-noise ratios for the highest two altitude bands are too low to allow measurement of He or HD + ^{3}He. The latitude is indicated on the lower axis, and closest approach is denoted by the vertical dotted lines. The spacecraft altitudes above the 1-bar pressure level are shown by the gray curves. The spacecraft moves from right to left in each plot. The corotation longitude and rev number are shown in the upper left and upper right corner of each plot, respectively. The final plunge into the atmosphere is labeled; the telemetry from the spacecraft ended at ~1360 km above the 1-bar pressure level.

the peak distributions of masses 28 u and 15 u shift into alignment with the closest-approach latitude near 5°S. The lowest-altitude band shows symmetric distributions of masses 15 u and 28 u. Their variation with latitude and altitude indicates that masses 15 u and 28 u are dominated by diffusive interaction with the H_2 atmosphere and are free-falling at a terminal velocity, dictated by the increasing hydrogen density and its slowing of the gas. Thus, as the altitude decreases and the background atmosphere increases in density, the maximum diffusion velocity decreases and the infalling material density increases to maintain flux continuity (apart from small effects at this altitude due to chemical loss and latitudinal spreading). This process allows us to estimate the diffusion velocity and determine the material's influx rate from the measured densities (see below).

Figure 4D reinforces the interpretation that the observed compounds originated in the rings. It indicates that for mass 15 u (CH_3 from methane), mass 28 u (CO, N_2, and C_2H_4), mass 44 u (CO_2 and propane), and a surrogate heavier mass from nanograin fragments at 78 u (benzene), the relative abundances are highest at the highest altitudes and reach a near-constant relative abundance below 2500 km as the material is diffusively coupled into Saturn's atmosphere.

Although the solar local time does not vary, the Saturn longitude does change from one orbit to the next. The abundance and distribution of methane, one of the major influx volatiles, varies by a factor of 3 about the average abundance (Fig. 4D, lowest-altitude values). These orbit-to-orbit variations indicate a spatial or temporal change of the volatile source on time scales that are faster than the horizontal diffusion time scale. However, there is no simple variability associated with Saturn's corotational longitude. There is a tentative link to longitudinal variations in the D68 ringlet near the inner edge of the D ring, in which a series of bright clumps appeared in 2015 (*17*) (Fig. 5). An influx that is spatially concentrated near the clumps may in part explain the spatial and/or temporal variations.

Impact fragmentation

The inflowing ring material impacted the INMS antechamber at 29 to 31 km s^{-1} (five times the speeds for which INMS was designed), leading to uncertainty in both absolute abundances and identification of the molecular species. At 31 km s^{-1}, molecules and nanoparticles carry 5 eV of kinetic energy for each atomic mass unit, which in some cases would have been sufficient to dissociate incoming particles and molecules as they impacted the CSN antechamber. However, studies of surface-induced dissociation [e.g., (*18*)] indicate that only about 25% of the impact energy is converted to internal energy of the molecule in the collision process. For H_2 this translates into ~2.5 eV, which is below its dissociation energy of 4.75 eV. Laboratory studies of dissociation in molecule-surface collisions (*19*) suggest that the interaction of H_2 with passive surfaces
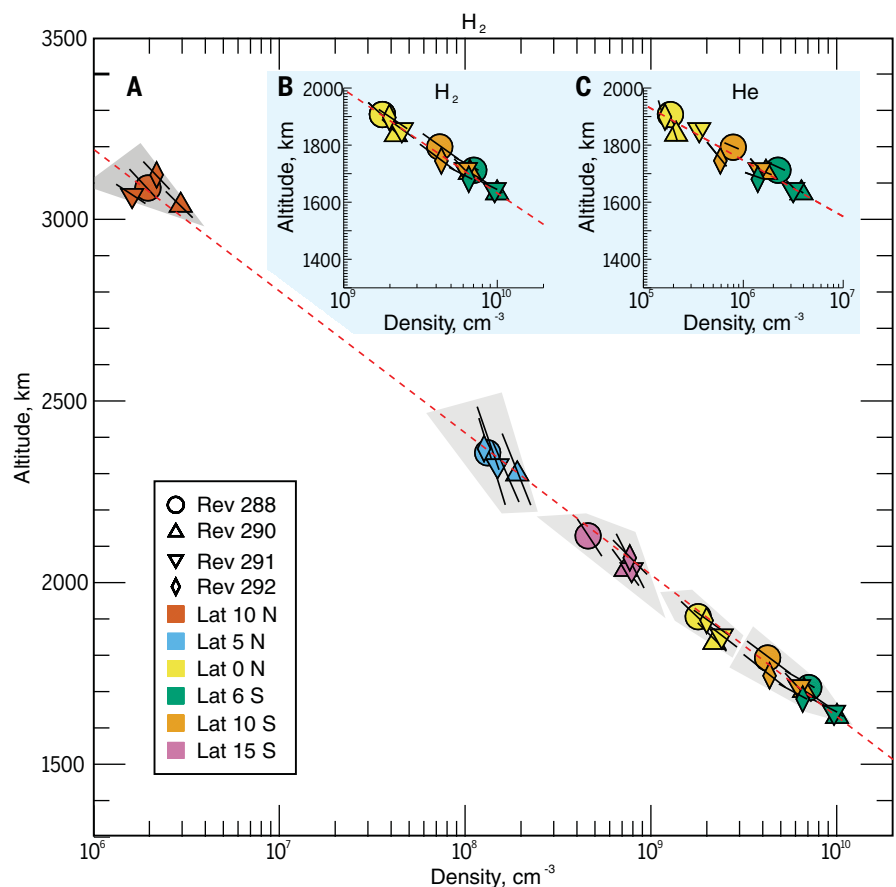
Fig. 3. Variation of H_2 and He with altitude and latitude. (**A**) H_2 densities (points) and locally determined scale heights (lines with 1σ uncertainty shaded in gray) for the altitude/latitude pairs at 10°N (orange), 5°N (light blue), 0° (yellow), closest approach (~5°S, green), and 10°S (brown) for revs 288, 290, 291, and 292 (denoted with different symbols). The common scale height (dashed red line) was determined by a least-squares fit to the density points; it is consistent with all the data except the scale heights at 5°N. (**B** and **C**) Zoomed versions include common scale heights of the lowest altitudes for H_2 (B) and He (C). The common scale heights (dashed red lines) show the 2:1 relation of H_2 to He expected for an isothermal atmosphere in diffusive equilibrium.

such as Ag proceeds as a direct interaction with surface atoms. These atomic interactions have a strong angular dependence, with a 10-eV neutral beam having a probability of dissociation of 50% for collinear collisions and >50% for perpendicular collisions. Interaction with a raw Ti surface, such as the CSN antechamber wall after fresh material is exposed through a grain impact, can lead to chemical adsorption (chemisorption) that produces metal hydrides (*20*). Chemisorption processes can also play a role with active metal surfaces or with oxygen-bearing compounds. Therefore, for H_2, CH_4, NH_3, and N_2, the effects of chemisorption in the fragmentation are likely moderate in our case, whereas for H_2O, CO, and CO_2, chemisorption/fragmentation will likely affect our measurements. For example, some of the CO observed at mass 28 u may be an impact fragment from CO_2, which does not affect our main conclusions. Impact fragmentation of CH_4 to produce CH_3 will likely result in reformation of CH_4 on the surface due to the high content and rapid movement of H radicals on the surface in

this hydrogen-rich environment. Similarly, CH_3 terminal groups from larger organics may also add to the methane signature, but because heavy organic compounds are at least an order of magnitude less abundant, this can increase the derived CH_4 value by at most 10%. Furthermore, this increase will be compensated by a decrease as a result of CH_4 impact fragmentation that forms CH and CH_2 fragments, which are lost from the antechamber before they can pick up two hydrogens in separate surface reactions because the surface coverage is relatively low (<30%). The net result is that we expect that the CH_4 statistical uncertainty encompasses the additional systematic uncertainties from various fragmentation processes.

The possibility of particle fragmentation inside the closed source does not affect our main conclusions about Saturn's atmospheric structure. With regard to H_2, the agreement of the outbound and inbound data for each orbit examined (290, 291, and 292) below 3000 km and the common scale heights observed at all altitudes, latitudes,

and orbits (Fig. 3A) indicate that the effects of dissociation are limited and do not affect the derived atmospheric structure below 3500 km. However, H_2 densities of 10^6 cm^{-3} in an extended outbound region above 4000 km have a constant slope (with density decreasing versus time and altitude) that tracks with the decay of water, which is known to adsorb to the instrument walls (21). This correlation suggests that grains have sputtered some raw Ti into the antechamber, allowing hydride formation on the antechamber surface. This hydrogen may have been subsequently displaced by the oxygen from water molecules (TiO and TiO$_2$ are more tightly bound than the metal hydride) clinging to the antechamber walls after the flyby, thereby creating a low level of H_2 production. This precludes the use of values for H_2 on the outbound phase above 4000 km, but does not affect our other results. However, the loss of water reacting with Ti vapor does suggest that the water abundance may be as
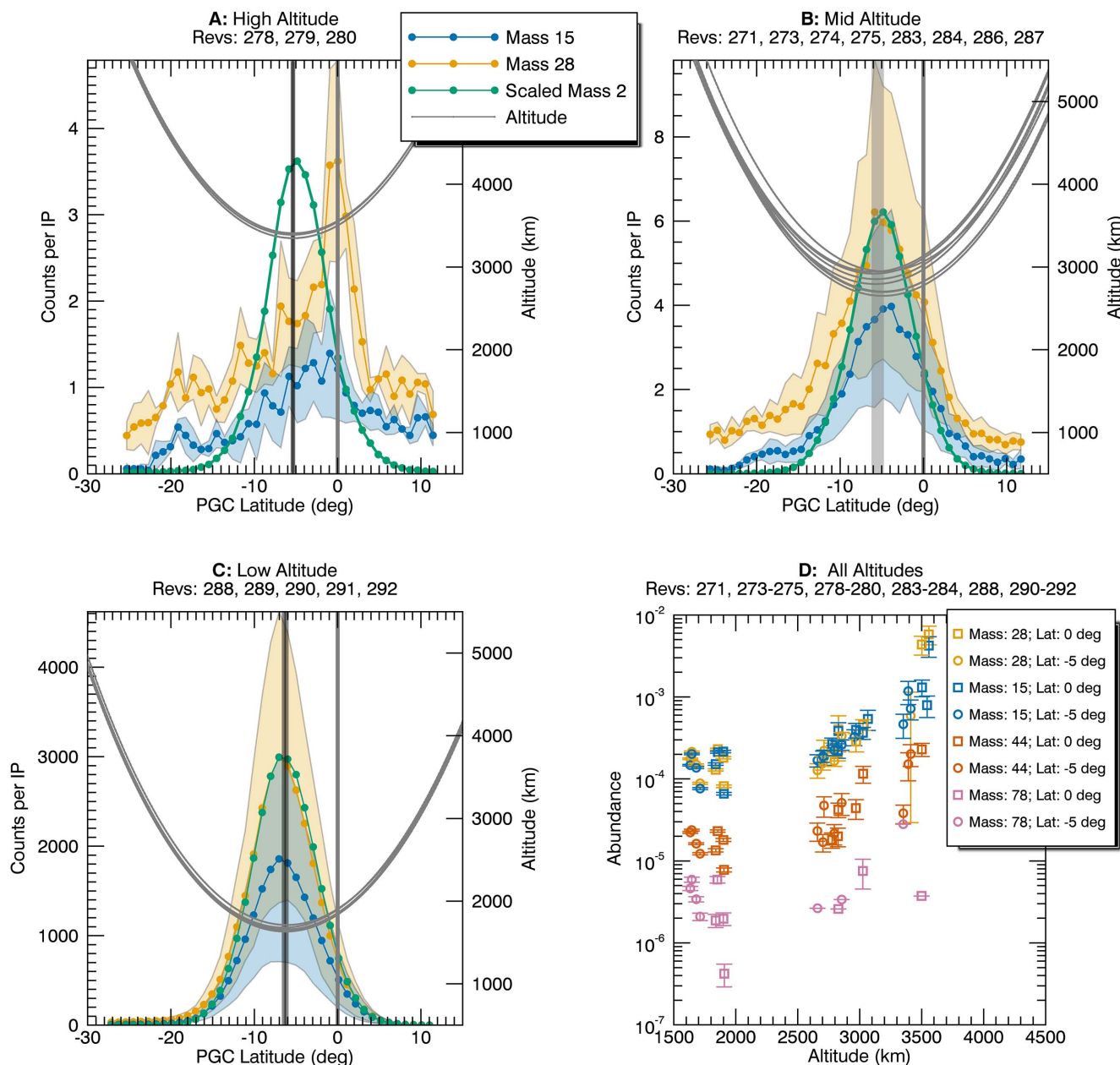
Fig. 4. Volatile inflow from the rings. (**A** to **C**) Plots of latitude and altitude relative to 1 bar for the 15 u and 28 u mass spectral features [in counts per 31-ms integration period (IP) with 1σ uncertainty as colored shading] averaged over each altitude band. The mass 2 u counts (scaled to match the 28 u peak value) are also provided for reference. The high-altitude band comprises revs 278 to 280 (A); the mid-altitude band, revs 271, 273 to 275, 283, 284, 286, and 287 (B); and the lowest-altitude band, revs 288 to 292 (C). The distributions shift in latitude from 0° to closest approach (~5°S) as altitudes decrease. The mass 28 u feature in the highest-altitude bin has both a broad latitude feature, like the 15 u feature, and a sharp spike at the equator indicating contributions from both volatiles (CO and/or N$_2$) and grains (C$_2$H$_4$ fragments from organic grain impacts). (**D**) Abundances (with 1σ error bars) for 28 u (carbon monoxide, molecular nitrogen, and ethylene), 15 u (methane), 44 u (carbon dioxide plus a propane impact fragment), and 78 u (a benzene impact fragment) as a function of altitude at the equator (squares) and at 5°S latitude (circles). The equatorial latitude is the location of ring plane influx, whereas 5°S latitude represents the closest approach to the atmosphere as a result of the flyby geometry (see Fig. 2).

much as 30% larger than reported in Table 1 and Fig. 6. Because the deuterium hydride has a lower activation barrier for formation than the H hydride (20), which led to H/D fractionation during the flyby, the present dataset is difficult to use for studying the H/D ratio of the atmosphere and rings.

The strongest evidence that CH_4 and N_2 are not simply by-products of fragmentation of heavy compounds, but instead native volatile gases, comes from the latitude distributions of mass 15 u and mass 28 u at the highest altitudes (3500 to 4000 km; see Fig. 4 and related text). The mass 15 u (methane) and mass 28 u signals both show a broad latitude distribution running from 20°S to 10°N latitude but differ in peak location; this finding constitutes evidence for a latitudinal spread of the native volatile compounds (CO and N_2), combined with C_2H_4 impact fragments of the organics nanograins from the peak near the ring plane. This leads us to conclude that methane, nitrogen, and carbon monoxide are native volatiles originating in the rings. However, we cannot rule out the possibility that some of the carbon monoxide is a fragment by-product of carbon dioxide, for which we also have evidence in the spectra. We have no corresponding evidence for the presence of water, carbon dioxide, or ammonia. They may be native volatiles or fragments from nanograins. Given our uncertainty of the fragmentation processes, it is fortunate that this rough classification between native volatiles and fragments derived from nanograins

has no effect on the mass inflow flux we derive below, although it does affect how these compounds react chemically with the atmosphere and ionosphere.

A plethora of organics

INMS spectra from Saturn's exosphere include signal over the full range of neutral masses, up to 99 u (Fig. 6). Species with mass exceeding 46 u are present in all six of the final low-altitude orbits, consistent with a local source for this material. The count rate distributions are highest around 16 u, 28 u, 44 u, 56 u, and 78 u, indicating an organic-rich spectrum. These spectra are more complex than predicted by models, with contributions from many different chemical compounds.

Our assessment of the composition of the inflowing material observed on revs 290, 291, and 292 gives the following fractions by weight: methane, 16 ± 3%; ammonia, 2.4 ± 0.5%; water, 24 ± 5%; molecular nitrogen and carbon monoxide (CO/N_2), 20 ± 3%; carbon dioxide, 0.5 ± 0.1%; and organic compounds, 37 ± 5%. The values reported are the mean of the orbits analyzed. To account for physical adsorption (physisorption) and chemisorption to the instrument walls (21), we generated integrated spectra (Fig. 6). For masses with a high tendency to interact with the walls of the antechamber (those in rev 290 with a ratio of outbound to inbound counts greater than 2 at 1750 km and with a maximum count rate greater than 40), the integrated spectra show the integrated signal at each mass over the full time

period for which the signal at 18 u is above the background level. The remaining masses are integrated using the time window from 500 s before to 500 s after closest approach. We used a standard fitting procedure (22). Using data from calibrations of the INMS engineering model and the National Institute of Standards and Technology mass spectral library (23) to determine the dissociative fragmentation patterns and absolute calibration, we constrain the abundances of carbon dioxide, CO/N_2, water, ammonia, methane, hydrogen, and helium for each integrated spectrum. The remaining counts at masses ≥12 u are attributed to organic species. Some signal may be due in part to inorganic S-bearing species such as H_2S or SO_2; however, the overall abundance of these compounds is consistently < 0.1% by mass of the inflowing material and less than our quoted uncertainties.

The ring particle composition (i.e., compounds other than H_2 and He) is approximately 37 weight percent (wt %) organic compounds heavier than CH_4. The other abundant ring particle compounds are water, CO/N_2, methane, ammonia, and carbon dioxide. Signals at 12 u and 14 u constrain the abundances of CO and N_2, respectively, and suggest that an inorganic component is likely present at 28 u. However, the value for CO reported is an upper limit, as the constraint from 12 u does not account for ionization fragments from organics or from impact fragmentation of CO_2 internal to the instrument. The organic fraction itself is fitted well by hydrocarbons, but the
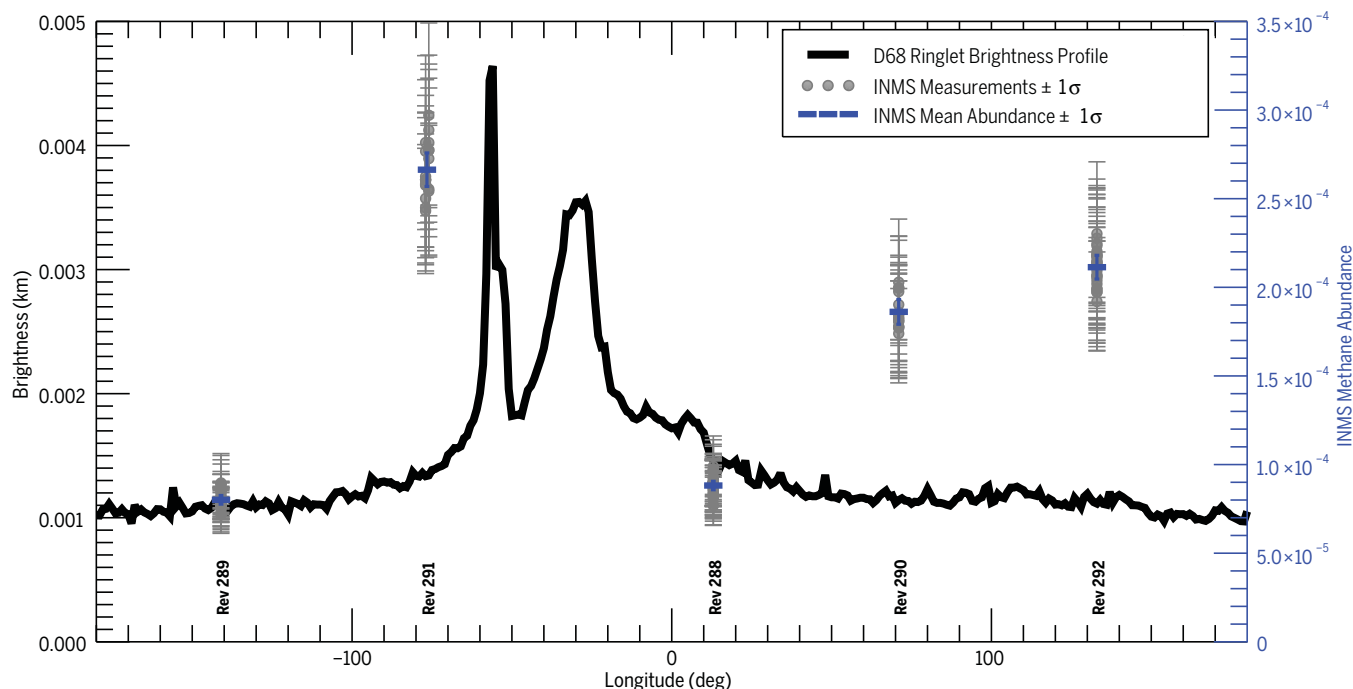
Fig. 5. Relationship between lower-altitude methane abundances and the reflected brightness of the D68 ringlet. The radially integrated brightness of the D68 ringlet (black line; see methods section) is plotted as a function of the Keplerian longitude associated with the clumps in the D68 ringlet, which has a radial distance of 67,630 km and an orbital period of 4.93 hours. Reported brightness is relative; the measurement is unitless but once integrated radially is provided in km. Overlain are the abundances (right axis) for methane observed by INMS between 1725 and 1750 km altitude on inbound in the low-altitude revs (288 to 292; gray points) and their mean at each longitude (blue rectangle with 1σ uncertainty). The highest abundances occur for rev 291, just behind the brightest parts of D68.
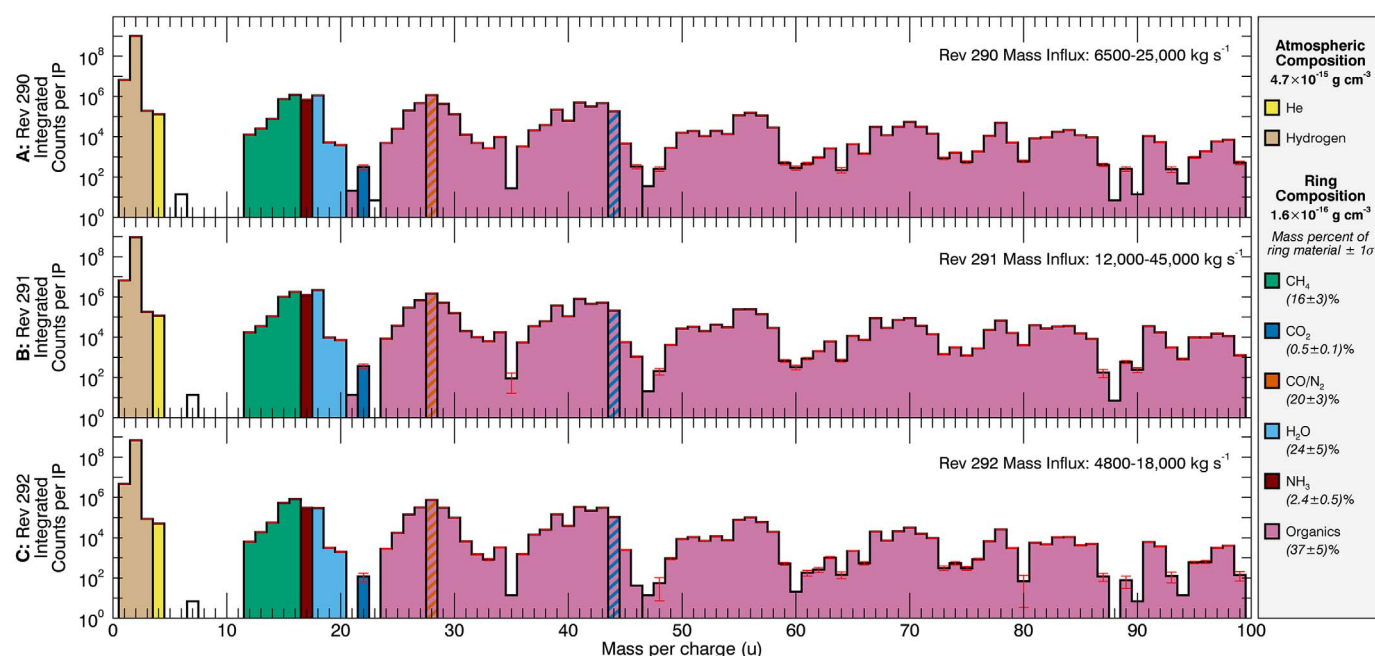
Fig. 6. Integrated mass spectra for the last three orbits of the Grand Finale. (A to C) Background-subtracted data integrated over the flyby are shown for revs 290, 291, and 292 along with 1σ error bars (red, in most cases narrower than plotting line). INMS has unitary mass resolution. The *y* axis is in counts per integration period (IP, 31 ms) integrated over the closest-approach data. The coloring corresponds to the dominant parent molecule for each mass. Masses 28 u and 44 u are both dominated by multiple components: CO/N_2 plus organics (28 u) and CO_2 plus organics (44 u). The side panel gives deconvolved mass density: H_2 and He from the atmosphere; and other compounds from the rings, whose percentage contributions to the ring material are also shown. The mass influx rate for each rev is shown at the top right of each panel. Mass influx rates and composition are summarized in Table 1.

presence of O- or N-bearing organics is not excluded. Water, 28 u, and methane are the most abundant volatiles with ratios relative to H_2 on the order of 10^{-4} for the integrated spectra. The spectra are also consistent with the presence of aromatic species, including the possible signal of benzene at 78 u. We estimate that the total organic mass density is on the order of 10^{-16} g cm^{-3}. The organic compounds identified may be either indigenous compounds present in the atmosphere or the products of high-speed impacts with INMS (see above). In either case, these spectra suggest abundant native organic material. The mass of hydrocarbons detected by INMS is equivalent to the mass of ~10^6 cm^{-3} nanoparticles with masses in the range detected by MIMI (*16*). Because MIMI detected far fewer particles, most of the mass measured by INMS appears to be below the 8000 u lower measurement limit of MIMI (*24*).

Masses above ~70 u are poorly fit by compounds with primary masses in the range of INMS. Because studies of heavy extraterrestrial organics indicate that these heavier organic compounds include aromatics [e.g., (*25*)], common aromatic molecules such as naphthalene were added to the spectral library as candidate parent species for this region. Compounds produced by ice irradiation experiments (*26, 27*) were also considered. We find that this mass region is consistent with the presence of aromatic compounds comprising ~5 to 10 wt % of organics. Aromatic material would be subject to impact dissociation

that may yield aliphatic compounds, so this abundance of aromatic material is a lower limit.

Estimating the mass influx in the equatorial region

We estimated the mass influx of material in the equatorial region by combining calculations of the material velocities and densities. From the mass deconvolution described above, we estimated the densities of methane, water, ammonia, nitrogen, carbon monoxide, carbon dioxide, and organic fragments measured at an altitude of 2000 km during revs 290, 291, and 292 (see Fig. 6). We used two different approaches to determine the appropriate downward diffusion velocity. One is based on the limiting flux equation (*28*):

$$v = \frac{D_{12}}{H}\left(1 - \frac{m_1}{m}\right) \qquad (1)$$

where H is the scale height (we use a value of 150 km based on the INMS data), m_1 is the mass of the minor species of interest, and m is the mean mass of the atmosphere (we use a value of 2 u). D_{12} is the coefficient for diffusion of species 1 in species 2. The alternative approach uses the diffusion coefficients calculated using a simple hydrostatic model adapted from previous work (*29*). The methane diffusion velocity at 2000 km is calculated using both methods for a range of influx values of methane. The maximum (1 × 10^4 cm s^{-1}) and minimum (4.5 × 10^3 cm s^{-1}) dif-

fusion velocities are then taken as bounding cases. The diffusion velocities for the other materials are scaled by the ratio to methane of their diffusion coefficients flowing through H_2 (*30*). The range of diffusion velocities used in this calculation is inclusive enough that assumptions about whether methane, ammonia, water, nitrogen, carbon monoxide, and carbon dioxide are present at 2000 km as volatiles or are derived as fragments of a larger organic moiety will have little effect on the diffusion velocity calculation. The equatorial latitude band used in the mass influx calculation is 8° wide, which is the half width of the low-altitude measurements in Fig. 4C. The difference between the maximum and minimum values is completely dominated by the uncertainty of determining the diffusion velocities, which is a systematic uncertainty that is constant from one orbit to the next. The uncertainty due to the densities of the components is ~1%, indicating that there is variation in the mass influx between rev 291 and the other two revs (290 and 292) we analyzed. However, the mass fraction of a given component is not statistically different from orbit to orbit. Calculated mass influx rates for revs 290, 291, and 292 and the average composition of material from Saturn's atmosphere and from ring influx are reported in Table 1 and Fig. 6.

Ionospheric measurements

The open source on the Cassini INMS was designed with the primary purpose of measuring

reactive neutrals and ambient ions in Titan's ionosphere and upper atmosphere (*11*), where flyby velocities were in the range of 6 km s^{-1}. However, the Cassini Grand Finale orbits resulted in spacecraft speeds relative to the atmosphere of ~31 km s^{-1}. At these speeds, INMS ion measurements were limited to <8 u. Despite the limited mass range, INMS measurements of light ions can be combined with Radio Plasma Wave Spectrometer (RPWS) measurements of the free electron content (*31*, *32*) to produce a more complete picture of the ionosphere. The INMS open source has a narrow field of view (< 2° cone) relative to the closed source (180° cone). Therefore, the spacecraft must be used to point the instrument into the ram direction to allow measurements of ions. This orientation occurred on a limited set of Grand Finale revs (283, 287, 288, and 292).

Figure 7 shows the ionospheric data obtained by INMS during the Cassini Grand Finale. Figure 7, A and B, shows the mid-altitude band (revs 283 and 287; refer to Fig. 1 for context). These time series indicate that for the altitudes and latitudes covered by these orbits, the free electron density is within a factor of 2 of the INMS measured light ion density. The asymmetry between north and south latitudes above 15° can be attributed to ring shadow effects during the autumnal solstice illumination (*33*). This reduces incoming solar ultraviolet flux, lowering the photoionization, although the effect is somewhat mitigated by the long lifetimes of the H$^+$ and H$_3^+$ ions and the associated transport effects. Figure 7, C and D, shows revs 288 and 292 from the lowest set of altitudes, where the near-equatorial ionosphere shows a large difference between the free electron density and the light ion density, indicating the presence of heavier ions (>8 u) that must account for the bulk of the ionospheric density (more than 75% of the ions are >8 u). This heavy-ion region is asymmetric with respect to the equator, ranging in latitude from ~2°N to 12°S. However, it does roughly correlate with the inflow of volatiles and organic nanograins from the D ring discussed above, and closely matches MIMI Charge Energy Mass Spectrometer (CHEMS) measurements of nanoparticles (*16*). The molecular volatiles (water, methane, ammonia, carbon dioxide) can easily convert the long-lived protons of the ionosphere into molecular ions with shorter lifetimes, decreasing the overall electron density (*6*), as described below.

Figure 7E shows additional measurements of the ionosphere for rev 288. The individual light ion concentrations for H$^+$, H$_2^+$, H$_3^+$, and He$^+$ are indicated. The H$^+$ and H$_3^+$ densities decrease in the near-equatorial heavy-ion region, consistent with an inflow of heavier molecules. The H$_2^+$ ions measured are the primary product of ionization in the Saturn ionosphere and are created by ionization of H$_2$, the most abundant neutral, by solar extreme ultraviolet radiation (*34*). They are a good tracer of the ionization process. Figure 7E also shows the scaled value of the positive nanoparticles measured by CHEMS (*16*) added to the INMS light ion density. The second-order

latitudinal structure of the CHEMS measurements of positive nanoparticles is closely correlated with the free electron density of the bulk ionosphere. Although the scaling factor is ~10^6 and these very large positive ions cannot themselves account for the secondary structure, they represent the positive member of a dusty plasma that contains both neutral nanoparticles and negative-ion nanoparticles, which appears to affect the recombination of the primary positive molecular ions that dominate the equatorial ionosphere.
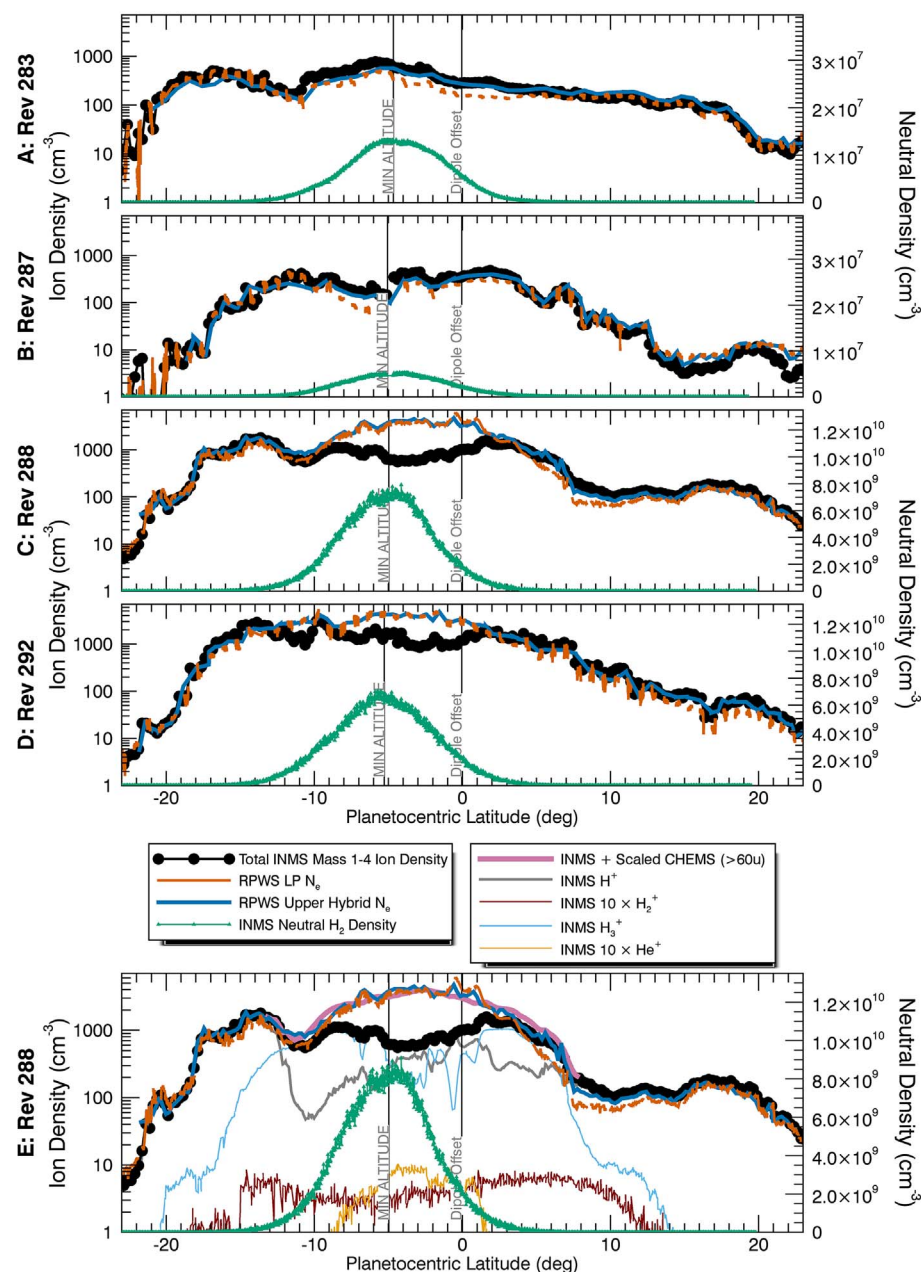
Fig. 7. INMS ionospheric observations during the Grand Finale. (**A** to **D**) Summary data for all four proximal orbits in which INMS obtained ionospheric data. The sum of the light ion densities for masses 1 to 4 u is shown by a black line. The RPWS electron density using the Langmuir Probe (LP) (*31*) is shown in dark orange and using a hybrid method (*32*) with the thick blue line (both on the left axis). The atmospheric H$_2$ density is shown by a green line (right axis). (**E**) The same data as (C) with the addition of the INMS-measured H$^+$, H$_3^+$, H$_2^+$, and He$^+$ ion densities (left axis) and the atmospheric H$_2$ density (right axis). The INMS ion density values are uncertain by about 21% at the ionospheric peak, rising to 25% at ±20° latitude; this is due to the combination of systematic uncertainty in the detector gain (*14*) and the statistical uncertainty. Also shown is the MIMI CHEMS measured nanoparticle positive ion density (*16*) multiplied by 10^6 (pink line) and added to the INMS light ion density, which corresponds to the electron density structure.
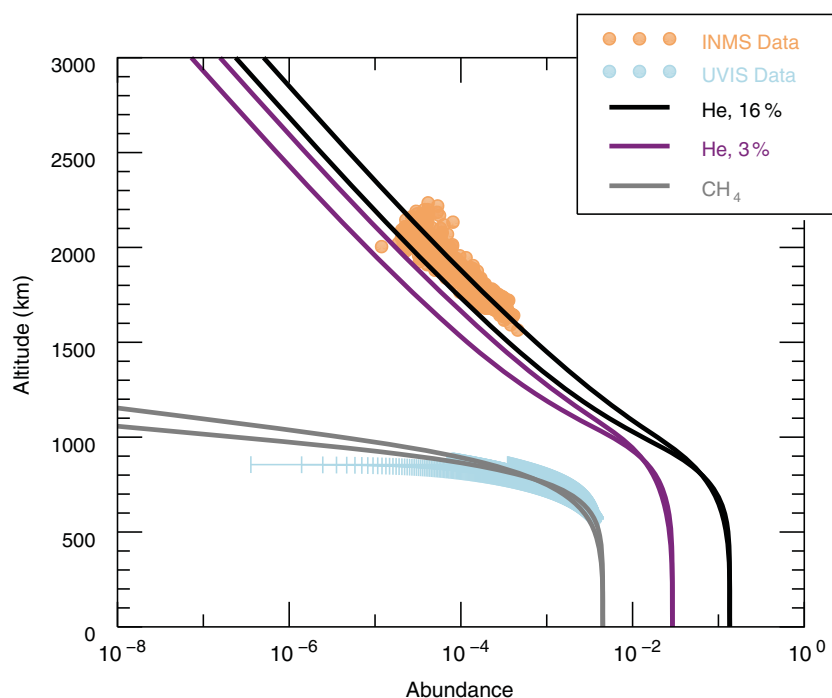
Fig. 8. Relationship of the INMS-measured He/H$_2$ ratio to the homosphere helium abundance. Altitude-dependent profiles for the abundances of methane (gray curves and light blue dots) and helium (black and purple lines and sandy brown dots) are shown. Methane data are from UVIS; He data are from INMS. Two scenarios for the homosphere He/H$_2$ ratio are shown: 0.03 in purple (*38*) and 0.16 in black (*35*), which is close to the jovian value (*4*). The solid and dash-dotted lines for each assumed abundance use two different formulations for the turbulent diffusion (eddy diffusion) coefficient (*29*, *35*).

Implications for Saturn's atmosphere and ionosphere
Atmospheric structure

The analyses shown in Figs. 2 and 3 indicate a background atmospheric structure that is consistent with predictions shown in Fig. 1 for both hydrogen and helium. However, species heavier than helium are far more abundant than predicted. The differences between the predicted and observed atmosphere are largely confined to the excess volatiles that we have concluded flow in from the D ring, as discussed below.

The measured hydrogen and helium abundances are compared to the models (*35*) in Fig. 8. The vertical profiles of the helium and methane abundances calculated by the hydrostatic model (*29*), which was used to benchmark the nonhydrostatic Global Ionosphere-Thermosphere Model (*36*, *37*), are shown. Methane abundances for altitudes below those observed by INMS are taken from the combination of Cassini Ultraviolet Imaging Spectrograph (UVIS) and Composite Infrared Spectrometer (CIRS) data (*35*). As can be seen in Fig. 8, the relatively large uncertainties in the methane abundance allow both approaches to reproduce the UVIS/CIRS methane data equivalently well, even though they use widely different versions of the eddy diffusion coefficient.

Also shown in Fig. 8 are two different scenarios for the deep-atmosphere He/H$_2$ ratio. The purple curves adopt a He/H$_2$ ratio of 0.03, consistent with a helium abundance of ~0.0291 reported from the Voyager measurements (*38*). We use the ratio of 0.03 to represent the most likely lower bound for the well-mixed atmosphere value for helium (*35*). The other set of helium curves represent a He/H$_2$ ratio in the well-mixed lower-altitude region of the atmosphere of ~0.16 (an abundance of helium of ~0.1355); this value is more consistent with the recent analysis (*35*), which reports a homosphere abundance of ~0.11 ± 0.02 inferred from UVIS and CIRS data. The latter approach brackets the helium abundances obtained directly from the INMS measurements, whereas the Voyager-derived curves systematically fall below the data (Fig. 8). This comparison suggests that the INMS helium data are more consistent with a nearly jovian homosphere abundance of ~0.1355 (*39*). In contrast, the model that uses a lower He/H$_2$ ratio systematically fails to reproduce the INMS data (*15*).

These results for the homospheric ratio of He/H$_2$ have implications for understanding the internal structure and evolution of Saturn. The conventional explanation for the excess infrared luminosity of Saturn relative to the expected thermal emission is that cooling over time leads to the demixing of helium from hydrogen, with the heavier helium raining out into the deeper interior and generating heat (*40*). Our measurements and modeling permit a modest depletion of helium but are inconsistent with a strong depletion relative to the protosolar He/H$_2$ ratio of 0.19 (*41*). As an example, the nominal INMS value for the He/H$_2$ ratio is 0.16 (Fig. 8), which is similar to the value at Jupiter (0.157 ± 0.003) (*42*). However, some additional helium rain in Saturn beyond that in Jupiter is allowed, as the INMS data are consistent with a He/H$_2$ ratio as low as 0.10. A range between 0.10 and 0.16 would maintain the viability of helium rain—a process that is consistent with He/H$_2$ < 0.12 in the well-mixed atmosphere (*43*)—as the cause of excess luminosity. This range is also consistent with the most recently derived He/H$_2$ ratio of 0.11 to 0.16 from Voyager (*44*) and with the Cassini UVIS-CIRS value of 0.09 to 0.13 (*35*). Overall, the homospheric helium abundance from INMS may be slightly higher than previous estimates, but the uncertainties are large.

Ionospheric structure

The presence of the light-ion species observed by INMS in the ionosphere (H$^+$, H$_2^+$, H$_3^+$, He$^+$) was predicted almost 40 years ago by a model of a neutral atmosphere dominated by H$_2$ and He (*1*). Ion and neutral measurements made by INMS in the ionosphere are consistent in that they both indicate the presence of an additional heavy molecular species (both neutral and ionized) in the equatorial upper atmosphere. Water-group neutral and ion species were predicted, but the present Cassini data indicate that the chemical composition of the material falling inward from the rings is concentrated at the equator, is chemically much more complex than predicted, and includes a substantial organic component, perhaps in the form of nanoparticles.

Dissociative and nondissociative photoionization of molecular hydrogen (and to a lesser degree He) by solar extreme ultraviolet radiation is the source of ionization in the equatorial ionosphere. The H$_2^+$ and H$^+$ ions thus produced undergo a series of ion-neutral reactions, generating other ion species such as H$_3^+$ via the fast reaction H$_2^+$ + H$_2$ → H$_3^+$ + H. The H$_2^+$ production rate along the spacecraft track can be determined empirically by multiplying the measured H$_2^+$ density by the measured H$_2$ density (both shown for rev 288 in Fig. 7E) and by the rate coefficient of 2 × 10^{-9} cm^3 s^{-1} (*34*). The production rate is ~8 cm^{-3} s^{-1} near closest approach. At these altitudes, the effects of approximately 50% opacity in the extreme ultraviolet are evident in the production rate, indicating that the spacecraft's closest approach nearly reached the altitude of peak ion production. This effect is also evident as the dip near closest approach in the H$_2^+$ densities in Fig. 7.

Figure 7E also shows a broad gap near closest approach between the total light-ion densities measured by INMS and the electron densities measured by RPWS. Assuming quasi-neutrality (that is, the ion density approximately equals the electron density) and neglecting negative ions, this suggests the existence of an ion with a mass beyond the upper limit of the open source for these orbits (8 u). This ion, or collection of ions, is more abundant than light ions in the

main ionospheric layer. To maintain consistency with the neutral composition provided by the closed source, the heavy ion cannot be solely a water-derived molecular ion. Simple ion chemistry for the light ions can put limits on the abundance of the heavy neutral ring ions with a large dissociative recombination rate coefficient. The low H^+ and H_3^+ densities measured near closest approach require fast reactions with a molecular volatile at an abundance of approximately 10^{-4} (*34*), consistent with the INMS neutral data. This simple ion analysis does not indicate the identity of the molecular volatile, which likely includes methane, ammonia, water, and carbon dioxide as measured by INMS. These compositional changes are discussed in (*45*).

Simple photochemistry for a single major ion (i.e., H_2^+) states that the total ion production rate equals the ion-electron loss rate from dissociative recombination. This leads to an expression for the electron density N_e:

$$N_e = \left(\frac{P_{H_2^+}}{\alpha}\right)^{1/2} \quad (2)$$

where $P_{H_2^+}$ is the H_2^+ production rate and $\alpha \approx 5 \times 10^{-7}$ cm^3 s^{-1} is a typical dissociative recombination rate coefficient (*46*). Using the peak production rate near closest approach from Fig. 7 yields a photochemical equilibrium value of $N_e \approx 10^4$ cm^{-3}, in good agreement with the peak electron density measured by RPWS (*31*, *32*). This agreement suggests that the role of negative ions and/or particles in determining the charge balance may be relatively minor.

It is well known in planetary and terrestrial aeronomy (*47*) that chemistry dominates at lower altitudes in an ionosphere, whereas transport processes become important at higher altitudes where the collision frequency is low. We expect this to be true for Saturn's ionosphere as well. The chemical lifetime of the major molecular compounds varies from ~200 s near closest approach up to ~2000 s (i.e., about 30 min) near the upper edge of the heavy-ion gap/layer. The H^+ chemical lifetime should be controlled by the abundance of the heavy neutral compounds, which increases rapidly with altitude, via ion-neutral reactions. Plasma transport is largely constrained to proceed along the magnetic field (*47*), which in the equatorial region is almost horizontal. The H_2^+ production rate sharply falls off by a factor of ~10 in the shadow of the planet's B ring (Fig. 7E), and the H_3^+ density also falls off rapidly. However, the H^+ density, which is equivalent to N_e in this region, falls more slowly. This suggests that H^+ is not in chemical equilibrium in the altitude region near 2000 km and above, but that H^+ plasma is produced outside this region and flows into the shadowed area.

He^+ ions (4 u) are created by the photoionization of atmospheric He, which falls with altitude more rapidly than H_2 because of diffusive separation. Figure 7E shows that the He^+ density decreases more rapidly with altitude than does the H_2^+ density. However, some of the 4 u signal is contributed by H_2D^+.

Origin of volatiles in the thermosphere

Molecular hydrogen and helium, which are sourced from the well-mixed atmosphere via diffusive transport, are the most abundant neutral species in the upper atmosphere of Saturn. The next most abundant category of neutrals by mass is organics, followed by water, mass 28 u inorganics (CO and N_2), and methane (Fig. 6). Methane is too heavy to diffuse upward from the homosphere, so the source of methane must be external (see above). The source of CH_4 seems to be Saturn's rings. One possibility is that an icy carrier of CH_4 (e.g., clathrate hydrate) may be present inside the ring particles, which volatilizes when heated by sunlight or ablation in Saturn's thermosphere. Any CH_4 gas released would diffuse into Saturn's atmosphere under gravity.

The volatile composition observed by INMS appears to be similar to material found in comets (*48*). This could be explained if Saturn's rings were formed from unprocessed primordial ices, derived from a thermally primitive precursor body such as a small icy moon. Alternatively, the similarity may be coincidental if species such as CH_4, NH_3, and CO are major products of the thermal/ultraviolet degradation of complex organics in an H_2-rich environment.

The mass of Saturn's C ring is ~10^{18} kg (*49*), about 0.03 times the mass of Saturn's moon Mimas. Therefore, if we use the mass influx inferred from the INMS measurements (4800 to 45,000 kg s^{-1}), we calculate a lifetime of 700,000 to 7 million years for the C ring. Yet this only reflects today's influx. The current influx is directly from the D ring rather than the C ring, which must be the ultimate supplier because the mass of the D ring [likely no more than 1% of the C ring mass (*50*)] can maintain current loss rates for only 7000 to 66,000 years—a very short amount of time in terms of solar history. It is unclear whether the C ring can lose 1% of its mass into the D ring by viscous spreading over that time period (*51*).

Although viscous spreading of the C ring is likely not the cause of mass transfer to the D ring (*51*), occasional transfer of ~1% of the mass of the C ring into the D ring region via a large ring-tilting event is feasible. These ring-tilting events involve a stream of planet-orbiting rubble crossing the ring plane somewhere in the C or D rings. The C ring provides the ultimate source, containing enough mass to last (at current influx rates) about 5% of the time that the rings themselves have existed (~200 million years) (*52*–*54*). The D ring could be repopulated sporadically by large impact events such as those that tilted the D and C ring plane (*55*). Once enough small particles are brought into the D ring region, exospheric drag would quickly drain them into the planet, as observed by Cassini.

We conjecture that one or more transient events occurred in the recent past that disturbed the D ring, or changed its mass and particle size distributions so that tenuous gas drag can more quickly cause it to fall into the planet. The latest such event appears to have been the one that perturbed the D68 ringlet (*17*). However, the

weak correlation shown in Fig. 5 is not compelling. Evidence of ionospheric depletion observed well before the formation of bright clumps in the D68 ringlet suggests that the material inflow may have been taking place for a longer period of time (*2*). Therefore, we examined the evidence for D ring perturbations over a longer time scale.

The D ring structure of irregularly spaced bands or belts has changed markedly since the Voyager flyby (*56*). A pattern within the ring has been interpreted as a vertical spiral ripple, possibly the result of ongoing wrapping by orbital evolution of particles in the initially tilted ring. From the wavelength of the wrap, which shortens with time, the event was dated to the early 1980s (*56*). The ripple was later found to extend through the C ring (*55*), indicating that the tilt was imposed by an impacting stream of rubble, perhaps from a disrupted comet 1 to 10 km in size with an extended node crossing the D and C rings. The event may have had two parts, separated by months (*57*). There is also evidence in Voyager data for two other disturbances that occurred in 1979 (*58*). All this sporadic disruption could plausibly have altered the properties of the D ring in such a way that today's flux (and/or today's D ring) are not necessarily typical of the last 100 million years. We expect these events to have happened at the same average rate going backward, producing variability in the D ring.

Our results for the delivery of ring materials have implications for the composition of Saturn's deep atmosphere (stratosphere and troposphere). Previous modeling (*7*) has suggested that the delivery of oxygen, in the form of water, can explain the presence of CO seen in the stratosphere. In addition to water, we observe carbon monoxide and carbon dioxide influx that can contribute to the oxygen inventory (Fig. 6).

Because of the influx of CH_4 and other sources of carbon, Saturn may have acquired an apparent methane enrichment (i.e., higher C/H ratio) relative to the protosolar value. Observations using the CIRS instrument onboard Cassini indicate that Saturn's methane enrichment over the protosolar value is 2.25 ± 0.55 times the enrichment seen in Jupiter (*59*). Our INMS observations indicate an influx of methane between 3×10^{28} and 2×10^{29} molecules per second entering the equatorial atmosphere, and 2.5 times as much mass in the form of other organics. If we assume that inflowing methane spreads over the globe, this is equivalent to an influx of 7×10^{11} to 4.8×10^{12} m^{-2} s^{-1} throughout the atmosphere. By calculating the column density of methane in the thermosphere, stratosphere, and troposphere above an altitude of 50 km, where the contribution function of CIRS peaked (*59*), we can estimate how long the observed methane influx from the rings would need to be sustained to raise the enrichment to 2.25 times that of Jupiter. The estimated time is ~7 million to 110 million years—within approximately an order of magnitude of the estimated lifetime of the rings themselves (see above). This slow buildup occurs in the stratosphere and troposphere, because the predominant methane flux at this altitude is

from the deep interior produced by recycled heavier hydrocarbons that photochemically formed in the stratosphere and later diffused down into the interior. Thus, the methane flux from above provides a very slow shift in the steady-state concentration that builds up over time in the stratosphere and troposphere. The organic carbon nanograin material, with a mass influx 2.5 times that of methane, could be chemically recycled deep in the atmosphere to increase the methane content in the deeper interior and drive a larger interior outflux of methane. However, our derived infalling material composition (Fig. 6) includes influx of NH_3, and it is unclear whether prolonged, continuous delivery of ring-derived NH_3 would be consistent with existing upper limits on the $^{15}N/^{14}N$ ratio in Saturn ([60]).

Conclusions

The Cassini INMS measured in situ the atmospheric and ionospheric composition of Saturn's equatorial atmosphere during a series of flybys between the atmosphere and the D ring in the Grand Finale phase of the mission. Water, methane, ammonia, carbon monoxide and/or molecular nitrogen, and carbon dioxide enter Saturn's atmosphere from the D ring along the ring plane. This influx is expected to affect the equatorial ionospheric chemistry by converting the H^+ and H_3^+ ions into heavier molecular ions, producing a depletion of ionospheric density previously observed in radio occultation observations ([10]). However, this may not explain the full extent of small-scale electron depletions observed by other Cassini instruments ([33]). INMS data include evidence for an influx of organic-rich nanoparticles that further modifies the composition and structure of the equatorial ionosphere and may circulate throughout the low- and mid-latitude thermosphere. Over long time scales, this infalling material may affect the carbon and oxygen content of the observed atmosphere.

Methods

The data on the D68 ringlet brightness distribution (Fig. 5) are from a sequence of images obtained by the Imaging Science Subsystem (ISS) onboard the Cassini spacecraft on day 229 of 2017, during rev 289. All images were calibrated using the standard CISSCAL routines, which remove dark currents, apply flat-field corrections, and convert the observed brightness data to I/F, a standardized measure of reflectance that is unity for a surface illuminated and viewed at normal incidence ([62], [63]). These calibrated images were geometrically transformed with the appropriate SPICE kernels ([64]), and the pointing was refined on the basis of the observed locations of stars in the field of view ([27]). For each image, the brightness data were then reprojected onto regular grids of radii and inertial longitude (i.e., longitude measured relative to the ascending node of the rings in the J2000 coordinate system). Each column of the reprojected maps then provides a radial profile of D68 at a single inertial longitude. Radial profiles were co-added to generate longitudinal brightness profiles. Because

the ring material orbits the planet, these profiles are constructed in a corotating longitude system with an assumed mean motion of 1751.7° per day and a reference epoch time of 300000000 TDB (Barycentric Dynamical Time), which is 2009-185T17:18:54 UTC (Coordinated Universal Time).

The Cassini images did not have sufficient spatial resolution to discern D68's internal structure, so the ringlet's brightness is quantified in terms of its equivalent width (EW), which is the radially integrated I/F of the ringlet over the radius range 67,550 to 67,700 km above a background level given by a linear fit to the signal levels on either side of the ringlet (67,000 to 67,500 km and 67,750 to 68,250 km, respectively). The estimates of the ringlet's equivalent width are converted to normal equivalent width (NEW) by multiplying the EW values by the cosine of the ring's emission angle. For features with low optical depth such as D68, NEW is independent of ring opening angle.

REFERENCES AND NOTES

1. J. H. Waite Jr., S. K. Atreya, A. F. Nagy, The ionosphere of Saturn: Predictions for Pioneer 11. *Geophys. Res. Lett.* **6**, 723–726 (1979). doi: 10.1029/GL006i009p00723

2. A. J. Kliore et al., Structure of the ionosphere and atmosphere of Saturn From Pioneer 11 Saturn radio occultation. *J. Geophys. Res.* **85**, 5857–5870 (1980). doi: 10.1029/JA085iA11p05857

3. G. L. Tyler et al., Radio science with Voyager 2 at Saturn: Atmosphere and ionosphere and the masses of Mimas, Tethys, and Iapetus. *Science* **215**, 553–558 (1982). doi: 10.1126/science.215.4532.553; pmid: 17771277

4. M. L. Kaiser, M. D. Desch, J. E. P. Connerney, Saturn's ionosphere: Inferred electron densities. *J. Geophys. Res.* **89**, 2371–2376 (1984). doi: 10.1029/JA089iA04p02371

5. J. E. P. Connerney, Magnetic connection for Saturn's rings and atmosphere. *Geophys. Res. Lett.* **13**, 773–776 (1986). doi: 10.1029/GL013i008p00773

6. J. E. P. Connerney, J. H. Waite, New model of Saturn's ionosphere with an influx of water from the rings. *Nature* **312**, 136–138 (1984). doi: 10.1038/312136a0

7. J. I. Moses, S. F. Bass, The effects of external material on the chemistry and structure of Saturn's ionosphere. *J. Geophys. Res.* **105**, 7013–7052 (2000). doi: 10.1029/1999JE001172

8. L. Moore, I. Mueller-Wodarg, M. Galand, A. Kliore, M. Mendillo, Latitudinal variations in Saturn's ionosphere: Cassini measurements and model comparisons. *J. Geophys. Res.* **115**, A11317 (2010). doi: 10.1029/2010JA015692

9. J. O'Donoghue et al., The domination of Saturn's low-latitude ionosphere by ring 'rain'. *Nature* **496**, 193–195 (2013). doi: 10.1038/nature12049; pmid: 23579676

10. A. J. Kliore et al., The ionosphere of Saturn as observed by the Cassini Radio Science System. *Geophys. Res. Lett.* **41**, 5778–5782 (2014). doi: 10.1002/2014GL060512

11. J. H. Waite Jr. et al., The Cassini Ion and Neutral Mass Spectrometer (INMS) Investigation. *Space Sci. Rev.* **114**, 113–231 (2004). doi: 10.1007/s11214-004-1408-2

12. I. C. F. Müller-Wodarg, L. Moore, M. Galand, S. Miller, M. Mendillo, Magnetosphere–atmosphere coupling at Saturn: 1 – Response of thermosphere and ionosphere to steady state polar forcing. *Icarus* **221**, 481–494 (2012). doi: 10.1016/j.icarus.2012.08.034

13. T. T. Koskinen et al., Saturn's variable thermosphere from Cassini/UVIS occultations. *Icarus* **260**, 174–189 (2015). doi: 10.1016/j.icarus.2015.07.008

14. B. D. Teolis et al., A Revised Sensitivity Model for Cassini INMS: Results at Titan. *Space Sci. Rev.* **190**, 47–84 (2015). doi: 10.1007/s11214-014-0133-8

15. R. V. Yelle et al., Thermal Structure and Composition of Saturn's Upper Atmosphere from Cassini/INMS Measurements. *Geophys. Res. Lett.* 10.1029/2018GL078454 (2018).

16. D. Mitchell et al., Dust grains fall from Saturn's D-ring into its equatorial upper atmosphere. *Science* **362**, eaat2236 (2018).

17. M. M. Hedman, J. A. Burt, J. A. Burns, M. R. Showalter, Non-circular features in Saturn's D ring. D68. *Icarus* **233**, 147–162 (2014). doi: 10.1016/j.icarus.2014.01.022

18. J. de Maaijer-Gielbert, Á. Somogyi, V. H. Wysocki, P. G. Kistemaker, T. L. Weeding, Surface-induced dissociation of diphenyl ether. *Int. J. Mass Spectrom.* **174**, 81–94 (1998). doi: 10.1016/S0168-1176(97)00292-9

19. A. W. Kleyn, Dissociation in molecule-surface collisions. *J. Phys. Condens. Matter* **4**, 8375–8394 (1992). doi: 10.1088/0953-8984/4/44/003

20. A. Forni, G. Wiesenekker, E. J. Baerends, G. F. Tantardini, A dynamical study of the chemisorption of molecular hydrogen on the Cu(111) surface. *J. Phys. Condens. Matter* **7**, 7195–7207 (1995). doi: 10.1088/0953-8984/7/36/009

21. B. Teolis, M. E. Perry, B. A. Magee, J. Westlake, J. H. Waite, Detection and measurement of ice grains and gas distribution in the Enceladus plume by Cassini's Ion Neutral Mass Spectrometer. *J. Geophys. Res.* **115**, A09222 (2010). doi: 10.1029/2009JA015192

22. B. A. Magee et al., INMS-derived composition of Titan's upper atmosphere: Analysis methods and model comparison. *Planet. Space Sci.* **57**, 1895–1916 (2009). doi: 10.1016/j.pss.2009.06.016

23. S. E. Stein, "Mass Spectra," in *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*, P. J. Linstrom, W.G. Mallard, Eds. (National Institute of Standards and Technology); https://webbook.nist.gov/chemistry/.

24. M. E. Perry et al., Material flux from the rings of Saturn into its atmosphere. *Geophys. Res. Lett.* 10.1029/2018GL078575 (2018).

25. S. A. Sandford et al., Organics captured from comet 81P/Wild 2 by the Stardust spacecraft. *Science* **314**, 1720–1724 (2006). doi: 10.1126/science.1135841; pmid: 17170291

26. C. K. Materese, D. P. Cruikshank, S. A. Sandford, H. Imanaka, M. Nuevo, Ice chemistry on outer solar system bodies: Electron radiolysis of N_2-, CH_4-, and CO-containing ices. *Astrophys. J.* **812**, 150 (2015). doi: 10.1088/0004-637X/812/2/150

27. R. L. Hudson, M. E. Palumo, G. Strazzulla, M. H. Moore, J. F. Cooper, S. J. Sturner, in *The Solar System Beyond Neptune*, M. A. Barucci, H. Boehnhardt, D. P. Cruikshank, A. Morbidelli, Eds. (Univ. of Arizona Press, 2008), pp. 507–523.

28. P. M. Banks, G. Kockarts, *Aeronomy* (Academic Press, 1973).

29. J. M. Bell et al., Developing a self-consistent description of Titan's upper atmosphere without hydrodynamic escape. *J. Geophys. Res.* **119**, 4957–4972 (2014). doi: 10.1002/2014JA019781

30. B. G. Higgins, H. Binous, "Binary Diffusion Coefficients for Gases" (Wolfram Demonstrations Project, 2013); http://demonstrations.wolfram.com/BinaryDiffusionCoefficientsForGases.

31. L. Z. Hadid et al., Saturn's ionosphere: Electron density altitude profiles and D ring interaction from the Cassini Grand Finale. *Geophys. Res. Lett.* 10.1029/2018GL078004 (2018).

32. A. M. Persoon et al., Electron density distributions in Saturn's ionosphere. *Geophys. Res. Lett.* 10.1029/2018GL078020 (2018).

33. J.-E. Wahlund et al., In situ measurements of Saturn's ionosphere show that it is dynamic and interacts with the rings. *Science* **359**, 66–68 (2018). doi: 10.1126/science.aao4134; pmid: 29229651

34. T. E. Cravens, L. Moore, J. H. Waite Jr., R. Perryman, M. Perry, J.-E. Wahlund, A. Persoon, W. S. Kurth, The Ion Composition of Saturn's Equatorial Ionosphere as Observed by Cassini. *Geophys. Res. Lett.* (2018). doi: 10.1029/2018GL077868

35. T. Koskinen, S. Guerlet, Atmospheric structure and helium abundance on Saturn from Cassini/UVIS and CIRS observations. *Icarus* **307**, 161–171 (2018). doi: 10.1016/j.icarus.2018.02.020

36. J. M. Bell et al., Simulating the one-dimensional structure of Titan's upper atmosphere: 1. Formulation of the Titan Global Ionosphere-Thermosphere Model and benchmark simulations. *J. Geophys. Res.* **115**, E12002 (2010). doi: 10.1029/2010JE003636

37. A. J. Ridley, Y. Deng, G. Tóth, The global ionosphere–thermosphere model. *J. Atmos. Sol. Terr. Phys.* **68**, 839–864 (2006). doi: 10.1016/j.jastp.2006.01.008

38. B. J. Conrath, D. Gautier, R. A. Hanel, J. S. Hornstein, The helium abundance of Saturn from Voyager measurements. *Astrophys. J.* **282**, 807–815 (1984). doi: 10.1086/162267

39. L. Ben-Jaffel, I. Abbes, Helium abundance in giant planets and the local interstellar medium. *J. Phys. Conf. Ser.* **577**, 012003 (2015). doi: 10.1088/1742-6596/577/1/012003

40. D. J. Stevenson, E. E. Salpeter, The dynamics and helium distribution in hydrogen-helium fluid planets. *Astrophys. J. Suppl. Ser.* **35**, 239–261 (1977). doi: 10.1086/190479

41. M. Asplund, N. Grevesse, A. J. Sauval, P. Scott, The Chemical Composition of the Sun. *Annu. Rev. Astron. Astrophys.* **47**, 481–522 (2009). doi: 10.1146/annurev.astro.46.060407.145222

42. U. von Zahn, D. M. Hunten, G. Lehmacher, Helium in Jupiter's atmosphere: Results from the Galileo probe Helium Interferometer Experiment. *J. Geophys. Res.* **103**, 22815–22829 (1998). doi: 10.1029/98JE00695

43. J. J. Fortney, W. B. Hubbard, Phase separation in giant planets: Inhomogeneous evolution of Saturn. *Icarus* **164**, 228–243 (2003). doi: 10.1016/S0019-1035(03)00130-1

44. B. J. Conrath, D. Gautier, Saturn Helium Abundance: A Reanalysis of Voyager Measurements. *Icarus* **144**, 124–134 (2000). doi: 10.1006/icar.1999.6265

45. L. Moore et al., Models of Saturn's equatorial ionosphere based on in situ data from Cassini's Grand Finale. *Geophys. Res. Lett.* 10.1029/2018GL078162 (2018).

46. D. R. Bates, Dissociative recombination of polyatomic ions. *J. Phys. B* **24**, 3267–3284 (1991). doi: 10.1088/0953-4075/24/14/019

47. R. W. Schunk, A. F. Nagy, *Ionospheres: Physics, Plasma Physics, and Chemistry* (Cambridge Univ. Press, ed. 2, 2009).

48. D. Bockelée-Morvan, An overview of comet composition. *Proc. IAU* **7**, 261–274 (2011). doi: 10.1017/S1743921311025038

49. M. M. Hedman, P. D. Nicholson, More Kronoseismology with Saturn's rings. *Mon. Not. R. Astron. Soc.* **444**, 1369–1388 (2014). doi: 10.1093/mnras/stu1503

50. M. M. Hedman, F. Postberg, D. P. Hamilton, S. Renner, H.-W. Hsu, in *Planetary Ring Systems*, M. Tiscareno, C. Murray, Eds. (Cambridge Univ. Press, 2018), pp. 308–337.

51. Scaling the A ring viscosity v from (*61*) by the ratio of C ring to A ring optical depths, and assuming a spreading time $t = D^{2/\nu}$, where v is viscosity and D is the width of the C ring, the spreading time is comparable to the age of the Solar System, much longer than the currently believed age of the rings.

52. M. M. Hedman, P. D. Nicholson, The B-ring's surface mass density from hidden density waves: Less than meets the eye? *Icarus* **279**, 109–124 (2016). doi: 10.1016/j.icarus.2016.01.007

53. J. N. Cuzzi, P. R. Estrada, Compositional Evolution of Saturn's Rings Due to Meteoroid Bombardment. *Icarus* **132**, 1–35 (1998). doi: 10.1006/icar.1997.5863

54. S. Charnoz, L. Dones, L. W. Esposito, P. R. Estrada, M. M. Hedman, in *Saturn from Cassini-Huygens*, M. K. Dougherty, L. W. Esposito, S. M. Krimigis, Eds. (Springer, 2009), pp. 537–575.

55. M. M. Hedman, J. A. Burns, M. W. Evans, M. S. Tiscareno, C. C. Porco, Saturn's curiously corrugated C ring. *Science* **332**, 708–711 (2011). doi: 10.1126/science.1202238; pmid: 21454753

56. M. M. Hedman et al., Saturn's dynamic D ring. *Icarus* **188**, 89–107 (2007). doi: 10.1016/j.icarus.2006.11.017

57. M. M. Hedman, M. R. Showalter, Corrugations and eccentric spirals in Saturn's D ring: New insights into what happened at Saturn in 1983. *Icarus* **248**, 137–161 (2015). doi: 10.1016/j.icarus.2014.10.021

58. M. M. Hedman, M. R. Showalter, A new pattern in Saturn's D ring created in late 2011. *Icarus* **279**, 155–165 (2016). doi: 10.1016/j.icarus.2015.09.017

59. L. N. Fletcher, G. S. Orton, N. A. Teanby, P. G. J. Irwin, G. L. Bjoraker, Methane and its isotopologues on Saturn from Cassini/CIRS observations. *Icarus* **199**, 351–367 (2009). doi: 10.1016/j.icarus.2008.09.019

60. L. N. Fletcher et al., The origin of nitrogen on Jupiter and Saturn from the $^{15}N/^{14}N$ ratio. *Icarus* **238**, 170–190 (2014). doi: 10.1016/j.icarus.2014.05.007

61. J. E. Colwell, P. D. Nicholson, M. S. Tiscareno, C. D. Murray, R. G. French, E. A. Marouf, in *Saturn from Cassini-Huygens*, M. K. Dougherty, L. W. Esposito, S. M. Krimigis, Eds. (Springer, 2009), pp. 375–412.

62. C. C. Porco et al., Cassini Imaging Science: Instrument Characteristics and Anticipated Scientific Investigations at Saturn. *Space Sci. Rev.* **115**, 363–497 (2004). doi: 10.1007/s11214-004-1456-7

63. R. West et al., In-flight calibration of the Cassini imaging science sub-system cameras. *Planet. Space Sci.* **58**, 1475–1488 (2010). doi: 10.1016/j.pss.2010.07.006

64. C. H. Acton Jr., Ancillary data services of NASA's Navigation and Ancillary Information Facility. *Planet. Space Sci.* **44**, 65–70 (1996). doi: 10.1016/0032-0633(95)00107-7

ACKNOWLEDGMENTS

STRUCTURAL BIOLOGY

Two Patched molecules engage distinct sites on Hedgehog yielding a signaling-competent complex

Xiaofeng Qi, Philip Schmiege, Elias Coutavas, Xiaochun Li*

INTRODUCTION: Appropriate Hedgehog (HH) signaling is important for human health; excessive HH signaling leads to various cancers, whereas insufficient HH signaling results in birth defects. The mature N-terminal domain of the HH protein with N-terminal palmitate and C-terminal cholesterol modifications (referred to as "native HH-N" here) binds to its receptor, Patched-1 (PTCH1), to release its repression of the HH pathway, presumably by regulating the transport of a ligand that regulates the downstream oncoprotein Smoothened. The palmitate moiety of HH-N is essential for HH signaling; however, the mechanism of how the palmitate moiety of HH induces the signaling is not clear. The means by which HH protein recognizes PTCH1 is also controversial: A recent 1:1 PTCH1–HH-N complex structure visualized a palmitate-mediated binding site on HH-N, which was inconsistent with previous studies that implied a distinct, Ca^{2+}-mediated interface for PTCH1 and HH co-receptors.

RATIONALE: Cell biological studies indicated that mutations on either the palmitate-dominated interface or the Ca^{2+}-mediated interface of HH-N eliminated signaling, indicating that both HH-N interfaces are necessary for HH to inhibit PTCH1 activity. We managed to assemble a 2:1 PTCH1– HH-N complex using a biologically functional PTCH1 variant (PTCH1*) and native Sonic HH-N (SHH-N) in the presence of 1 mM $CaCl_2$ (a physiological extracellular calcium concentration) for structural investigation.

RESULTS: The cryo–electron microscopy (cryo-EM) structure of native SHH-N in complex with PTCH1* at 3.5-Å resolution demonstrates that one SHH-N molecule concomitantly engages both epitopes and binds two PTCH1 receptors (PTCH1-A and PTCH1-B) in an asymmetric manner. PTCH1-A and PTCH1-B exhibit several conformational differences and bind distinct portions of SHH-N. Structural comparisons suggest that the Ca^{2+}-mediated interface of SHH-N physiologically binds different HH binders, including PTCH1, HH co-receptors, and the SHH-N antagonist 5E1 antibody, to regulate HH signaling. Functional assays using PTCH1 or SHH-N mutants that disrupt the individual interfaces illustrate that simultaneous engagement of both interfaces between SHH-N and PTCH1 is essential for maximal signaling in cells.

Calculations using MOLE reveal a tunnel, ~150 Å in length with a minimum radius of 4 Å, that stretches through PTCH1-B but not PTCH1-A, which interacts with the palmitate of SHH-N. The path of this PTCH1-B tunnel includes the extracellular domain I (ECD-I), the cavity of the ECDs that accommodates the palmitate, and the transmembrane region. We introduced two point mutations (L427R and F1017E) on full-length human PTCH1 to block this tunnel by a ~3-Å salt bridge instead of by palmitate insertion. Cell biological assays show that unlike wild-type PTCH1, this variant was not able to repress the HH pathway; this finding supports our hypothesis that the palmitate can abolish the activity of PTCH1 by blocking the cavity of the ECDs.

CONCLUSION: Our structure of a signaling-competent complex consisting of SHH-N and an asymmetric arrangement of two PTCH1* molecules reconciles previous disparate findings of how HH-N interacts with PTCH1. The structural analysis reveals a tunnel in the PTCH1 molecule that can be blocked using mutations in the ECDs that mimic the palmitate insertion, thus abolishing HH pathway repression by PTCH1. Thus, our work illuminates how two sites on SHH-N serve to unite two PTCH1 receptors, thereby initiating a full HH signal. ∎
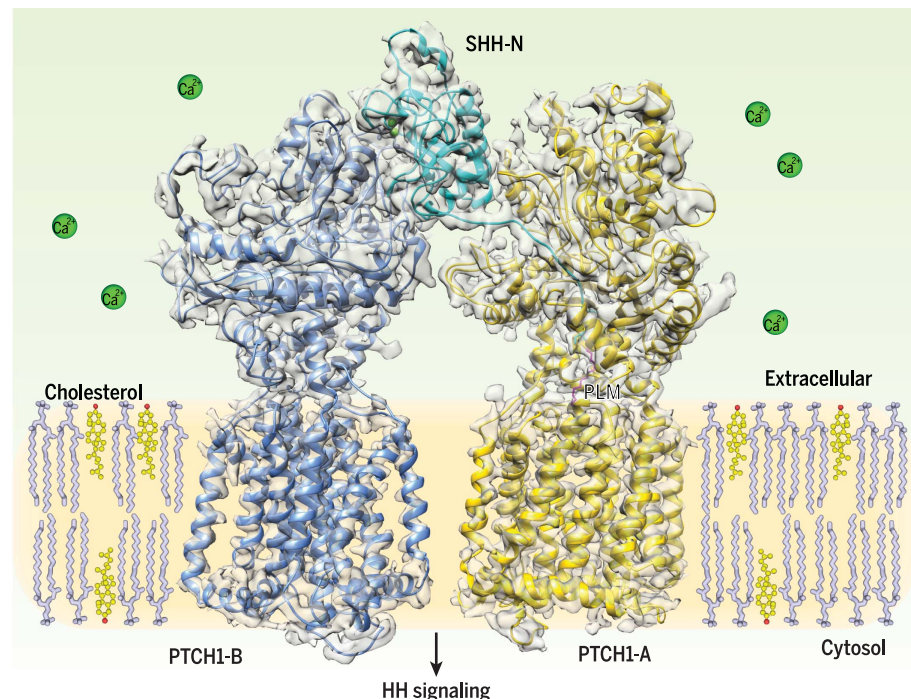
Cryo-EM structure of 2:1 human PTCH1*–SHH-N signaling-competent complex.
PTCH1-A, PTCH1-B, and SHH-N are colored yellow, light blue, and cyan, respectively, and are viewed from the side of the membrane. Palmitate (PLM) is shown as magenta sticks, calcium as green spheres. Cholesterol in the cell membrane is shown as ball-and-stick models colored yellow and red.

STRUCTURAL BIOLOGY

Two Patched molecules engage distinct sites on Hedgehog yielding a signaling-competent complex

Xiaofeng Qi[1], Philip Schmiege[1], Elias Coutavas[2], Xiaochun Li[1,3]*

Aberrant Hedgehog (HH) signaling leads to various types of cancer and birth defects. N-terminally palmitoylated HH initiates signaling by binding its receptor Patched-1 (PTCH1). A recent 1:1 PTCH1-HH complex structure visualized a palmitate-mediated binding site on HH, which was inconsistent with previous studies that implied a distinct, calcium-mediated binding site for PTCH1 and HH co-receptors. Our 3.5-angstrom resolution cryo–electron microscopy structure of native Sonic Hedgehog (SHH-N) in complex with PTCH1 at a physiological calcium concentration reconciles these disparate findings and demonstrates that one SHH-N molecule engages both epitopes to bind two PTCH1 receptors in an asymmetric manner. Functional assays using PTCH1 or SHH-N mutants that disrupt the individual interfaces illustrate that simultaneous engagement of both interfaces is required for efficient signaling in cells.

Appropriate Hedgehog (HH) signaling is important for human health: Excessive HH signaling leads to various cancers, whereas insufficient HH signaling results in birth defects (1–6). Nascent HH proteins enter the secretory pathway. After signal peptide cleavage, HH undergoes autocatalytic processing between its N- and C-terminal domains (HH-N and HH-C) in the endoplasmic reticulum (7). During this reaction, cholesterol is covalently linked to the C terminus of HH-N, which is necessary for its trafficking and distribution (8). Hedgehog acyltransferase subsequently adds palmitate to the α-amino group of the N-terminal cysteine of HH-N (9, 10). This modification is indispensable in activating the HH pathway (9–14). Activation occurs when the lipid-modified HH-N binds to its receptor, Patched-1 (PTCH1) protein (15). Notably, a palmitoylated SHH-N-terminal peptide bound to PTCH1 was found to partially stimulate HH signaling while the globular portion of HH could increase the binding to PTCH1 and facilitated PTCH1 export from cilia (16). In the absence of HH-N binding, PTCH1 represses the HH pathway presumably by regulating the transport of a ligand of its downstream protein, Smoothened (SMO) (2, 17, 18). Patched is a tumor suppressor involved in basal cell carcinoma and medulloblastoma (19), and SMO is an oncoprotein and a target of antitumor agents (20, 21).

We recently determined the cryo–electron microscopy (cryo-EM) structure of a functional human PTCH1 protein (PTCH1*) alone and in complex with Sonic Hedgehog (SHH-N) (22). Although PTCH1* lacks the internal loop between transmembrane helices 6 and 7 (TM6 and TM7) and the C-terminal cytoplasmic domain, it can still repress the HH pathway and be regulated by SHH-N, comparable to wild-type full-length PTCH1 in $Ptch1^{-/-}$ mouse embryo fibroblasts (MEFs) (22, 23). The structure shows that PTCH1* comprises two structurally homologous extracellular domains (ECDs) and 12 TMs. Notably, an endogenous ligand observed in the sterol-sensing domain (SSD), consisting of TMs 2 to 5, suggests that the SSD may play a role in the transport and/or regulation of the SMO ligand. The structure of the complex shows a palmitate-dominated interface between SHH-N and PTCH1 (22); this interface is opposite the interface of SHH-N with its co-receptors (24), which had been implicated in PTCH1 binding in previous studies (25–28) and can be regulated by calcium (24). Consistent with this, we reported biochemical assays showing that SHH-N lacking palmitate interacts with PTCH1 through the Ca^{2+}-mediated interface. However, unpalmitoylated SHH-N fails to stimulate HH signaling in the cell (16, 22).

Assembly of 2:1 PTCH1*–SHH-N complex

The extracellular Ca^{2+} concentration is typically in the range of 1 to 2 mM (29), and mutations on either the palmitate-dominated interface (22) or the Ca^{2+}-mediated interface (30) attenuated HH signaling, indicating that both SHH-N interfaces are necessary for maximal HH signaling in the cell. Previously, Ca^{2+} was not included in our purification buffer. To enable the formation of either or both interfaces, we mixed PTCH1* with native SHH-N containing its N-terminal palmitate and C-terminal cholesterol modifications in the presence of 1 mM $CaCl_2$. A 2:1 PTCH1*–SHH-N complex

was eluted ahead of the 1:1 complex by gel filtration and subjected to cryo-EM studies (Fig. 1A and fig. S1A).

We determined this structure at 3.5-Å resolution (Fig. 1B, fig. S1, B to D, and table S1). One SHH-N forms a complex with two PTCH1* molecules (PTCH1-A and PTCH1-B) using two distinct interfaces (Fig. 1C). The SHH-N palmitate inserts into the ECD cavity of PTCH1-A, and its Ca^{2+}-mediated interface directly binds the ECDs of PTCH1-B (Fig. 1C). PTCH1-A uses its ECD-I to bind the globular portion of SHH-N, while PTCH1-B engages both of its ECDs to bind the Ca^{2+}-mediated epitope (Fig. 1D). The interface between PTCH1-A and SHH-N is ~2300 Å² (including 350 Å² to accommodate the palmitate), whereas the interface between PTCH1-B and SHH-N is ~1100 Å². The transmembrane domains of the two PTCH1* molecules are arranged asymmetrically, with the TMs of PTCH1-B rotated about 150° relative to the TMs of PTCH1-A (Fig. 1E).

Interfaces between SHH-N and two PTCH1* molecules

The current map allows us to visualize the interactions between SHH-N and the two PTCH1* molecules (figs. S2 and S3). There are several hydrogen bonds and salt bridges between the N-terminal residues of SHH-N and the PTCH1-A ECDs, which form the palmitate-dominated interface. Only five residues of ECD-II (Asn[802], Tyr[804], Asn[940], Gln[973], and Tyr[1013]) interact with the N-terminal peptide (Np) and the palmitate of SHH-N; however, more than 10 residues from ECD-I contribute to this interaction (Fig. 2, A and B). A previous study showed that the mutation Pro[26] → Ala on SHH-N impairs signaling (16). Our structure reveals the molecular basis of this effect: The proline facilitates palmitate insertion into the hydrophobic cavity by keeping the N terminus of SHH-N in a bent position (Fig. 2C). In addition to the palmitate insertion, the palmitate-dominated interface includes α5 of PTCH1-A recognizing α1 and the N terminus of α2 of SHH-N (Fig. 2A). We previously showed that mutations on α5 of PTCH1 abolished PTCH1*–SHH-N binding and HH signaling (22).

The PTCH1-B molecule also uses its α5 helix and several loops of ECD-I to form the Ca^{2+}-mediated interface of SHH-N (Fig. 2, D and E). In addition, three residues (Glu[947], Asp[951], and Asp[954]) in an ECD-II loop (D loop) form salt bridges with Arg[153] and Arg[155] of SHH-N (Fig. 2F). The D-loop is not used in the PTCH1-A palmitate-mediated interaction. In the context of SHH-N without palmitate, mutation of Arg[153] causes loss of binding to PTCH1 (23). In the context of SHH-N with palmitate, the mutant also fails to stimulate HH signaling (30), consistent with the requirement for both interfaces. A structure of PTCH1 with unpalmitoylated SHH-N was recently published (31), which is consistent with our observation of the interface between PTCH1-B and native SHH-N.

PISA (proteins, interfaces, structures, and assemblies) server analysis (32) calculates that the $\Delta^i G$ between the palmitate and its dominated interface to PTCH1-A is –12.1 kcal/mol (P = 0.232),

[1]Department of Molecular Genetics, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA. [2]Laboratory of Cell Biology, The Rockefeller University, New York, NY 10065, USA. [3]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA.
*Corresponding author. Email: xiaochun.li@utsouthwestern.edu

more stable than the calcium-mediated interface to PTCH1-B ($\Delta^i G$ = –1.1 kcal/mol, P = 0.717). Overall, across both interfaces, residues in ECD-I contribute to more than 80% of the interactions with SHH-N. Residues Asp[217] and Glu[221] in helix α5 and Gly[378] in the loop connecting α3 and α3′ function in both interfaces; however, no residues in ECD-II are used in both of the two interfaces at the same time (Fig. 2G). Residues in ECD-II bind to the globular portion of SHH-N on the Ca^{2+}-mediated interface, and different residues in ECD-II are involved in the palmitate insertion. This distinction introduces an asymmetric arrangement of two PTCH1* molecules in the complex.

Superimposing PTCH1-A and SHH-N from the 2:1 complex with the PTCH1*–SHH-N 1:1 complex (PDB code 6D4J) reveals a slight change in the palmitate-dominated interface. SHH-N in the 2:1 complex has moved away from α5 of PTCH1-A as a result of the interaction of PTCH1-B with the Ca^{2+}-mediated interface (Fig. 2H). Intriguingly, there is a rod-like density of unknown identity among helices α5 to α7 of ECD-I in both PTCH1* molecules (fig. S4). This ligand may stabilize α5 of ECD-I in the conformation that binds SHH-N or could be a ligand that is regulated by PTCH1 in the HH pathway; however, this molecule and its precise function require identification and further analysis.

Structural comparison of PTCH1-A and PTCH1-B after SHH-N binding

PTCH1-A uses its ECD-I for binding the globular domain of SHH-N, and PTCH1-B uses a similar area of its ECD-I for SHH-N recruitment (Fig. 1C). To compare the conformations of the two PTCH1* molecules in the complex, we superimposed PTCH1-A and PTCH1-B: The α3 and α3′ helices of ECD-I in PTCH1-A are moved outward relative to PTCH1-B to accommodate the N terminus of SHH-N that inserts into the cavity of PTCH1-A (Fig. 3A). The Np and the palmitate of SHH-N trigger a loop in ECD-I to move, allowing insertion (Fig. 3A, red frame). The α5 in ECD-I is in distinct conformations in PTCH1-A and PTCH1-B because it is involved in different interfaces with SHH-N (Fig. 2, A and E, and Fig. 3A). The loop between α6 and α7 in ECD-I of PTCH1-B is shifted as a result of its interaction with Glu[89] in SHH-N (Figs. 2D and 3B). In the transmembrane domain, TM4 and TM12 have moved slightly in PTCH1-A relative to PTCH1-B (Fig. 3, C and D). There are some conformational changes between the two small α helices following TM1 and TM7 (pre-α1) and α4 of ECD-I, which form the cavity that accommodates the palmitate (Fig. 3C and fig. S3).

As seen in the 1:1 complex (22), two endogenous sterol-like ligand densities can be detected in the transmembrane domain of each PTCH1* molecule, one in the SSD and the other close to TM12 (figs. S2 and S5). The palmitate moiety in the cavity between the ECDs, together with TM4 and TM6, provides a hydrophobic environment for the endogenous ligand (Fig. 3D). The palmitate may affect the dynamics of the SSD by inserting into the center of the protein; a similar
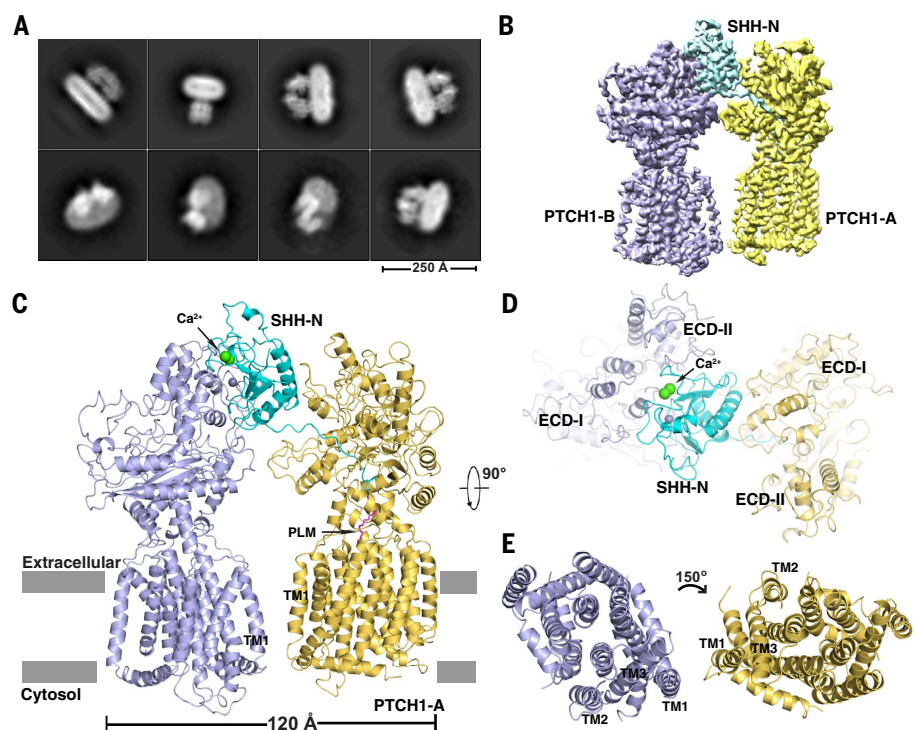
Fig. 1. Overall structure of 2:1 PTCH1*–SHH-N complex. (**A**) Cryo-EM 2D classification from RELION. (**B**) Cryo-EM map after Frealign refinement sharpened using BFACTOR.EXE with a resolution limit of 3.5 Å and a B-factor value of –100 Å². (**C**) Ribbon representation of the structure viewed from the side of the membrane. In this and succeeding figures, calcium and zinc are shown as spheres in green and gray, respectively; palmitate (PLM) is shown in magenta. (**D**) The ECDs (top view). (**E**) The transmembrane domains (top view).

mechanism was postulated to regulate cholesterol egress from the lysosome in the homologous NPC1 protein (33–35). The cavity between the ECDs might also serve as part of a tunnel for the transport of endogenous ligands from ECD-I or the TM domain (figs. S4 and S5). Blocking the tunnel by palmitate could then inactivate PTCH1.

Physiological importance of the Ca^{2+}-mediated interface of SHH-N

The palmitate of SHH-N facilitates an interaction with PTCH1 that is distinct from interactions with other HH-N partners (Fig. 4A). In contrast, the Ca^{2+}-mediated interface of SHH-N, which is required for PTCH1-B binding, has been reported to recruit other HH-N binders such as the co-receptors CDO and BOC (Fig. 4, B to F) (24). Consistent with this, previous experiments showed that the fibronectin type III (FnIII) domain of CDO competes with PTCH1 for HH-N binding in vitro (27). HH co-receptors (e.g., CDO and BOC) are essential for HH-mediated cell proliferation (30, 36): By binding the Ca^{2+}-mediated interface of HH-N, they may increase the local concentration of HH-N, which could promote the PTCH1–HH-N interaction (Fig. 4, C and D).

The commonly used monoclonal antibody 5E1 binds to the Ca^{2+}-mediated interface of SHH-N, blocking HH signaling (37, 38). Structural analysis indicates that the interface between 5E1 and SHH-N overlaps with that of PTCH1-B and SHH-N

(Fig. 4, B and E). However, 5E1 does not compete with the palmitate-dominated interface; we previously detected a PTCH1*–SHH-N–5E1 complex in pull-down assays (22). Hedgehog-interacting protein (HHIP) serves as a negative regulator of the HH pathway and blocks HH signaling (39) by directly binding the Ca^{2+}-mediated interface of SHH-N (28, 40) (Fig. 4F). Therefore, we speculate that the Ca^{2+}-mediated interface, rather than the palmitate-dominated interface, may be primarily regulated by the different HH binders in cells to prevent or promote the formation of the 2:1 complex, affecting signaling.

2:1 PTCH1–SHH-N complex in HH signaling

The structural analysis shows that most of the SHH-N surface is occupied by two PTCH1 molecules and that the N terminus of SHH-N is buried in the cavity formed by the ECDs of PTCH1-A (Fig. 1C). Interestingly, PTCH1-A and PTCH1-B use a similar area to recognize the globular portion of SHH-N, making it impossible for SHH-N to mediate the formation of a PTCH1 oligomer of higher order than a dimer. Our finding suggests that 2:1 PTCH1-HH complex could be the functional organization in HH signaling.

To investigate this further, we performed cell-based assays. SHH-light II cells are transfected with GLI-responsive firefly luciferase reporter and a constitutive renilla-luciferase expression

vector for the quantitative measurement of biologically active HH proteins. The palmitoylated SHH-N and its variants were collected from the medium and detected by an antibody to SHH. Using the SHH-N with either I111E/N115K or R153E mutations—located in the two interfaces in our 2:1 complex—substantially reduced signaling in cells (Fig. 5, A and B). The C24S palmitoylation site SHH-N mutant completely lost its

activity for HH signaling (Fig. 5B). Together, these studies show that both interfaces of SHH-N and its palmitate moiety are necessary for maximal HH signaling in cells and that the palmitate moiety is absolutely required.

Using $Ptch1^{-/-}$ MEFs with the same luciferase reporting system, we expressed full-length human PTCH1 with either N940R or E947R/D951R mutations on ECD-II to specifically block the

interaction on either side of SHH-N (Figs. 2G and 5A). Expression of either of the mutants alone repressed the HH pathway, and in neither case could the repression be released by introducing SHH-N (Fig. 5C). Additionally, our previous work showed that PTCH1 with a mutation of Glu[221], a residue that is necessary for the PTCH1–SHH-N interaction in both interfaces (Fig. 5A), can repress the HH pathway but does
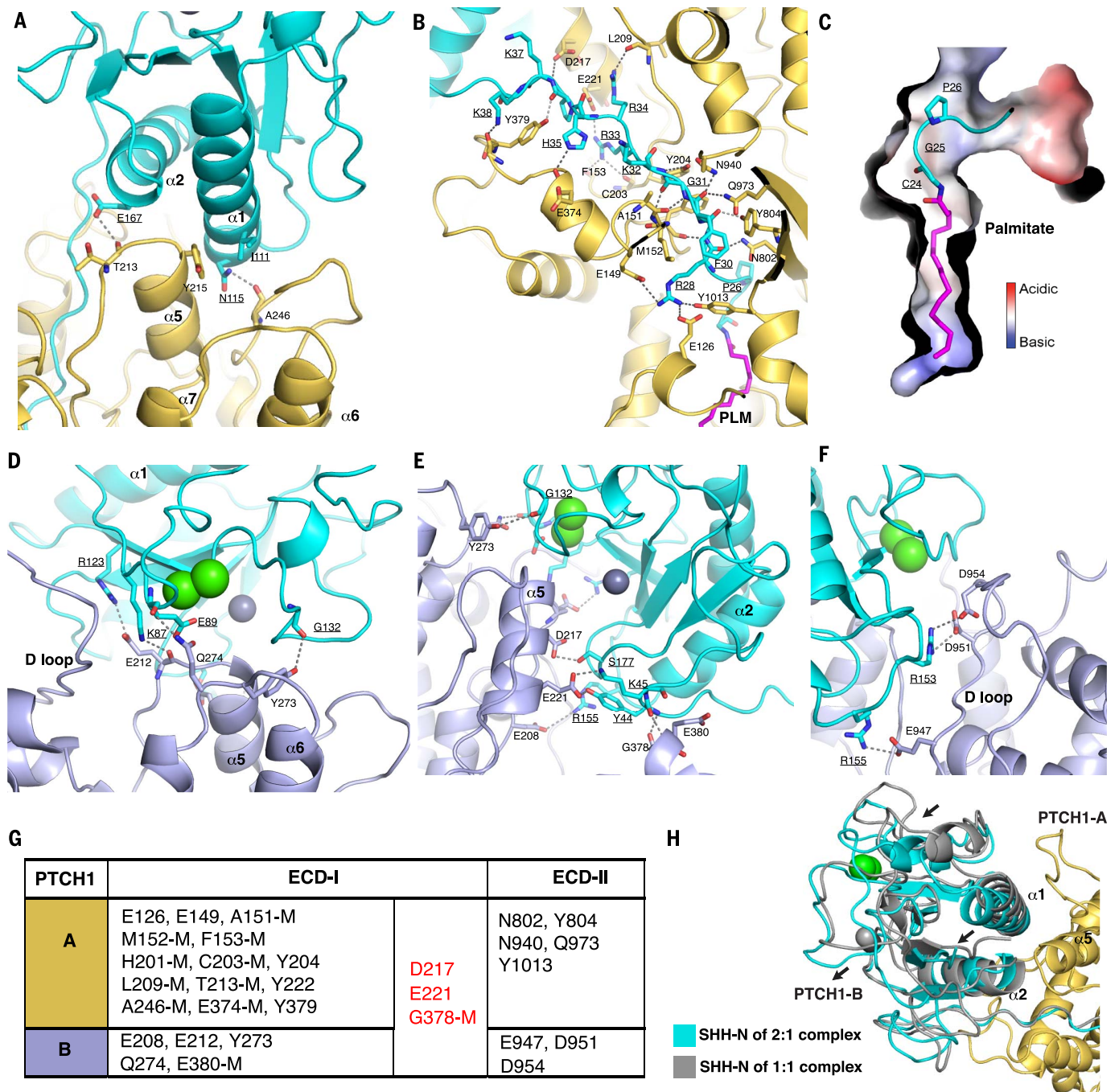
Fig. 2. The two interfaces of SHH-N bound to PTCH1*. (**A**) Details of α1 and α2 of SHH-N binding to PTCH1-A. (**B**) Details of the interaction between the N terminus of SHH-N and PTCH1-A. (**C**) The bend in the N terminus of SHH-N. (**D** to **F**) Interaction detail of the Ca^{2+}-mediated interface of SHH-N to PTCH1-B residues. (**G**) Summary of the residues of PTCH1*

bound to the distinct interfaces of SHH-N. M, main chain–involved interaction. (**H**) Structural comparison of SHH-N in 2:1 and 1:1 PTCH1*–SHH-N complexes. Amino acid residues: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr. Residues of SHH-N are underlined.

not respond to HH-N (*22*), supporting our current observations. Therefore, these assays validated our structure-based proposal that the 2:1 PTCH1–SHH-N complex is essential for maximal HH signaling.

The putative mechanism of palmitoylated HH–mediated signaling

The palmitate of SHH-N inserts into the cavity created by several hydrophobic residues of the ECDs, preventing any access of a putative ligand through the protein. MOLE calculations (*41*) show that there is a tunnel, ~150 Å in length with a minimum radius of 4 Å, that stretches throughout the PTCH1-B protein but not PTCH1-A, which interacts with the palmitate of SHH-N (Fig. 6A). The path of this tunnel includes the area where the unidentified ligand is located in ECD-I (fig. S4), plus the cavity of the ECDs that accommodates the palmitate and the transmembrane region. There are two gates of this tunnel in the transmembrane domain; one is near TM3 of the SSD and the other is near TM12. Notably, two unidentified sterol-like densities were detected in these two areas in the 2:1 PTCH1–SHH-N complex (fig. S5) as well as in the 1:1 PTCH1–SHH-N complex (*22*), implying that these gates are involved in the transport and/or regulation of the sterol-like molecules.

To test the function of the palmitate insertion, we introduced two point mutations (L427R and F1017E) on full-length human PTCH1 to block this tunnel by a ~3-Å salt bridge instead of by a palmitate insertion (Fig. 6B). This variant was transfected to *Ptch1*$^{-/-}$ MEFs, and unlike wild-type PTCH1, it was not able to repress the HH pathway (Fig. 6C). This finding supports our hypothesis that the palmitate could abolish PTCH1 repression of HH signaling by blocking the cavity of the ECDs. It also could explain how abolishing the palmitoylation of HH-N leads to embryonic defects in both *Drosophila* and mice (*10, 12, 13, 42–44*), which is a direct biological consequence of insufficient HH signal.

Before the HH-N recognizes PTCH1, co-receptors might capture the Ca^{2+}-mediated interface of HH-N by their flexible extracellular domains (*24*). This would increase the likelihood of forming a PTCH1–HH-N complex, a function reminiscent of the interleukin-2 receptor α-subunit (*45*). It is also possible for a PTCH1 molecule to engage the Ca^{2+}-mediated interface if the palmitoylated N terminus of HH-N is bound to other binders, such as various heparan sulfate proteoglycans (*46, 47*). This would allow PTCH1 to serve as a possible co-receptor to promote extracellular transport of HH-N. Because the globular portion of SHH-N triggers the internalization of PTCH1 (*16*), once HH-N recognizes PTCH1, the co-receptor–PTCH1–HH-N or 2:1 PTCH1–HH-N trimeric complex might be internalized, therefore creating a PTCH1-free environment that further promotes HH signaling (*22*).

Conclusions

Our structure of a signaling-competent complex consisting of SHH-N and an asymmetric arrange-
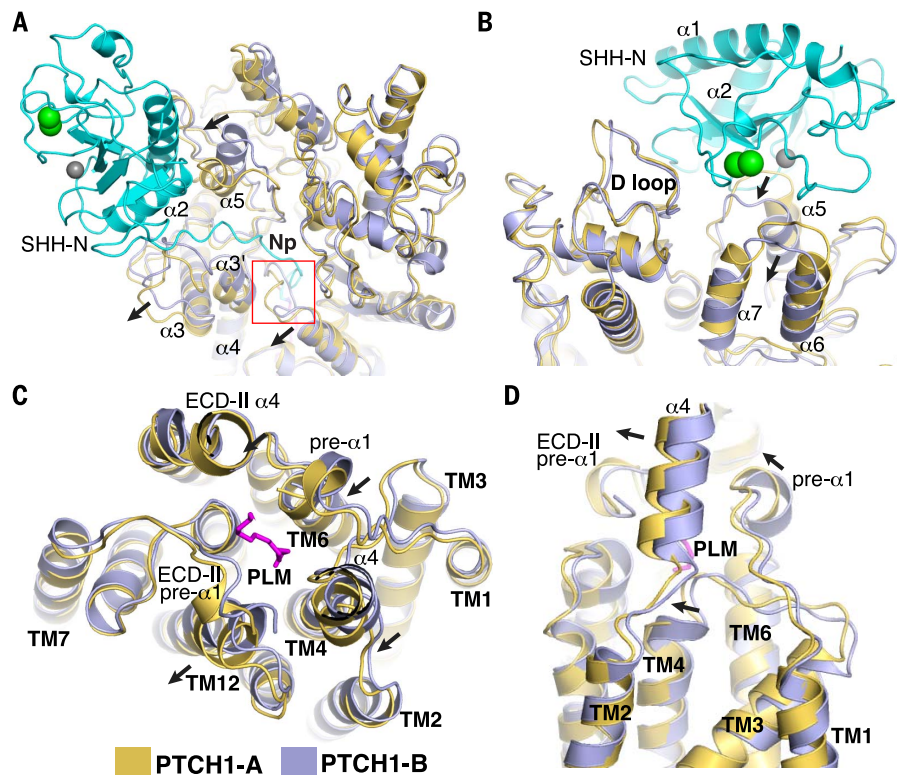
Fig. 3. Structural comparison of the two PTCH1* molecules. (**A**) Movement of structural elements of ECD-I in PTCH1-A after SHH-N insertion. (**B**) Superimposition of the ECDs of PTCH1* molecules (top view). (**C**) Superimposition of transmembrane domains of PTCH1* molecules (top view). (**D**) Superimposition of transmembrane domains of PTCH1* molecules from the side of the membrane.

ment of two PTCH1* molecules reconciles the disparate findings of how HH-N interacts with Patched. Our cell biological assays revealed that both PTCH1–SHH-N interactions are required for efficient signaling. The CTD and cytosolic loop between TM6 and TM7 have been removed from PTCH1*; therefore, further investigation is required to determine whether the complex results in any coupling between the cytosolic regions of full-length PTCH1.

The structural analysis reveals a tunnel in the PTCH1 molecule. Interestingly, the palmitate of SHH-N directly disrupts this tunnel when bound to PTCH1 (Fig. 6A). Mutations in the ECDs that mimic the palmitate insertion by blocking the ECD cavity abolish the repression of the HH pathway by PTCH1 (Fig. 6, B and C). Knowledge of the specific identity of the endogenous sterol-like molecules in ECD-I and the transmembrane region of PTCH1 would substantially facilitate our understanding of PTCH1 function and SMO activation.

Previous studies also show that HH-N co-receptors are necessary for HH signaling by binding to HH-N and forming a complex with PTCH1 (*30, 36, 48–50*). These HH-N co-receptors might play a role in increasing the local concentration of HH-N, thereby promoting PTCH1–HH-N interaction and HH signaling. The structure of PTCH1–HH-N in complex with its co-receptors will be informative in furthering

our understanding of their function in the HH pathway.

Materials and methods
Protein expression and purification

Human PTCH1* and SHH-N proteins were purified according to our previously published protocol (*22*). To assemble the PTCH1* and SHH-N complex, native SHH-N (purchased from R&D Systems, Cat # 8908-SH/CF) was mixed with purified PTCH1* at a 1:2 molar ratio and purified by size-exclusion chromatography using a Superdex-200 column (GE Healthcare) in buffer A (20 mM Hepes pH 7.5, 150 mM NaCl) with 0.06% Digitonin and 1 mM $CaCl_2$. The peak fractions were collected and concentrated to 5 mg/ml for grid preparation. The mutated and truncated DNA constructs were generated using Gibson Assembly (NEB) or QuikChange Mutagenesis kit (Agilent).

EM sample preparation and imaging

A freshly purified protein sample was added to Quantifoil R1.2/1.3 400 mesh Au holey carbon grids (Quantifoil), blotted using a Vitrobot Mark IV (FEI), and frozen in liquid ethane. The grids were imaged in a 300 keV Titan Krios (FEI) with a Gatan K2 Summit direct electron detector (Gatan). Data were collected at 1 Å per pixel with a dose rate of 8 electrons per physical pixel per second. Images were recorded for 10-s
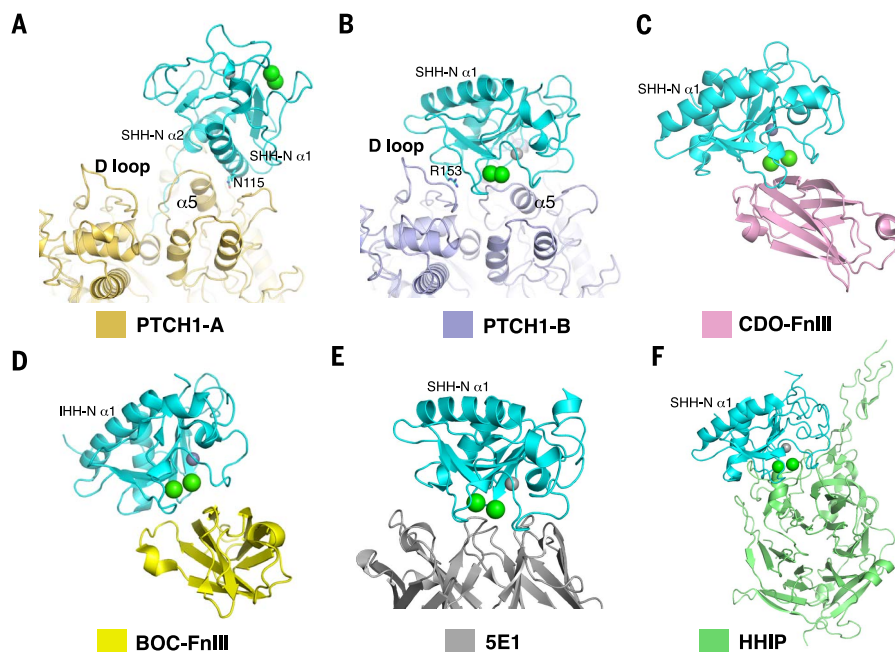
Fig. 4. Physiological importance of the Ca²⁺-mediated interface of SHH-N. (**A**) Interaction details of PTCH1-A and the globular portion of SHH-N. (**B**) Interaction details of PTCH1-B and SHH-N. (**C**) Structure of IHH-N with FnIII domain of CDO (PDB code 3D1M). (**D**) Structure of SHH-N with FnIII domain of BOC (PDB code 3N1M). (**E**) Structure of SHH-N with 5E1 Fab (PDB code 3MXW). (**F**) Structure of SHH-N with Hedgehog-interacting protein (HHIP) (PDB code 3HO5).

Fig. 5. Functional characterization of HH signal biogenesis. (**A**) Location of the key residues for HH signal biogenesis. The residues of SHH-N are underlined. (**B**) Mutation on either interface or palmitoylation site of SHH-N abolishes HH signaling. HH activity was measured by luciferase assay. Data are means ± SD (n = 3). SHH-N in conditioned medium was detected by Western blotting (inset). (**C**) HH signaling in $Ptch1^{-/-}$ MEFs transfected with PTCH1 and treated with SHH-N ligand. Luciferase activity was measured as described in the methods section. Data are means ± SD (n = 3). The expression of PTCH1 was detected by Western blotting (inset).

exposures in 50 subframes to give a total dose of 80 electrons per Å².

Image processing and 3D reconstruction

Dark-subtracted images were first normalized by gain reference that resulted in a pixel size of 1 Å per pixel. Drift correction was performed using the program MotionCorr (*51*). The contrast transfer function (CTF) was estimated using CTFFIND4 (*52*). To generate 2:1 PTCH1*–SHH-N templates for automatic picking, around 3000 particles were manually picked and classified by 2D classification in RELION (*53*). After auto-picking in RELION, the low-quality images and false-positive particles were removed manually. About 275k particles were extracted for subsequent 2D and 3D classification. We used the cryo-EM structure of human PTCH1* (EMD-7795) low-pass filtered to 60 Å as the initial model for 3D classification in RELION. The PTCH1* model of best class after 3D classification was used as the initial model for the final 3D classification in RELION. The best class, containing ~77k particles, provided a 5-Å map after 3D auto-refinement in RELION. Refinement was performed in FREALIGN (*54*) using this best class as the initial model. The global search was performed once without mask followed by another global search using the mask, which was generated using "relion_mask_create" with 6-Å extensions excluding the micelle. The final map is estimated to be 3.5 Å using the 0.143 cutoff criteria.

Model construction

To obtain better side-chain densities for model building, we sharpened the map using BFACTOR.EXE (author: N. Grigorieff) with a resolution limit of 3.5 Å and a B-factor value of –100 Å². The structure of human PTCH1*–SHH-N complex was docked to the map. The density of residues 1 to 75 (N-terminal domain), 608 to 618 and 721 to 730 (TM6-TM7 linker), and 1177 to 1188 (C terminus) of PTCH1* is not resolved nor built. Residues 890 to 915 of PTCH1* were built with polyalanine due to limited local resolution.

Model refinement and validation

The model was refined in real space using PHENIX (*55*) and also in reciprocal space using Refmac with secondary-structure restraints and stereochemical restraints (*56, 57*). Structure factors were calculated from a half-map (working) using the program SFall (*58*). Fourier shell correlations (FSCs) were calculated between the two half maps, the model against the working map, the other (free) half map, and full (sum) map (*59*). Local resolutions were estimated using Blocres (*60*). MolProbity (*61*) was used to validate the geometries of the model. Structure figures were generated using PyMOL (www.pymol.org) and Chimera (*62*).

HH reporter assays

Human SHH-N (24-197) was constructed into pcDNA3.1 vector with the signal sequence of human calreticulin at the N terminus as described (*16*). Secreted SHH-N was produced in HEK293 cells by transient transfection for 72 hours and

Fig. 6. Putative mechanism of palmitoylated HH–mediated signaling. (**A**) Structure of the 2:1 PTCH1*–SHH-N complex with the putative tunnel in blue; the two gates in the transmembrane domain are indicated. (**B**) Location of the residues for mimicking the palmitate insertion to block the ECDs' cavity. (**C**) HH signaling in $Ptch1^{-/-}$ MEFs transfected with PTCH1 and treated with SHH-N ligand. Luciferase activity was measured as described in the methods section. Data are means ± SD ($n = 3$). The expression of PTCH1 was detected by Western blotting (inset).

collected in DMEM, with 0.5% fetal bovine serum (FBS). SHH Light II cells, a stable cell line expressing firefly luciferase with an 8X-Gli promoter and Renilla luciferase with a constitutive promoter (from B. Chen and J. Kim), were used to measure HH pathway activity. SHH Light II cells were treated with the conditioned medium diluted in fresh DMEM with 0.5% newborn calf serum for 30 hours. To measure the activity of PTCH1 variants in HH signaling, the 8X-Gli-Firefly luciferase reporter transgene, a constitutive Renilla luciferase transgene, and a pcDNA3.1 vector encoding PTCH1 variants were transfected to $Ptch1^{-/-}$ MEFs (from B. Chen and J. Kim) using TransIT reagent (Mirus Bio LLC). After 24 hours, cells were serum-starved in DMEM with 0.5% FBS; 24 hours later, cells were treated with SHH-N conditioned medium for another 24 hours. Firefly and Renilla luciferase activity were measured using the Dual-Luciferase Reporter Assay System (Promega). All reporter assays were performed at least three times. The conditioned medium added was normalized based on Western blotting with anti-SHH antibody. The expression of PTCH1 variants and internal calnexin in MEF cells were detected by Western blotting with anti-PTCH1 antibody (GeneTex, 83771) and anti-calnexin antibody (Novus, NB100-1965), respectively. Each assay was reproduced at least three times.

REFERENCES AND NOTES

1. J. Taipale, P. A. Beachy, The Hedgehog and Wnt signalling pathways in cancer. *Nature* **411**, 349–354 (2001). doi: 10.1038/35077219; pmid: 11357142
2. P. W. Ingham, A. P. McMahon, Hedgehog signaling in animal development: Paradigms and principles. *Genes Dev.* **15**, 3059–3087 (2001). doi: 10.1101/gad.938601; pmid: 11731473
3. L. L. Rubin, F. J. de Sauvage, Targeting the Hedgehog pathway in cancer. *Nat. Rev. Drug Discov.* **5**, 1026–1033 (2006). doi: 10.1038/nrd2086; pmid: 17139287
4. J. Jiang, C. C. Hui, Hedgehog signaling in development and cancer. *Dev. Cell* **15**, 801–812 (2008). doi: 10.1016/j.devcel.2008.11.010; pmid: 19081070
5. J. Briscoe, P. P. Thérond, The mechanisms of Hedgehog signalling and its roles in development and disease. *Nat. Rev. Mol. Cell Biol.* **14**, 416–429 (2013). doi: 10.1038/nrm3598; pmid: 23719536
6. K. Petrov, B. M. Wierbowski, A. Salic, Sending and Receiving Hedgehog Signals. *Annu. Rev. Cell Dev. Biol.* **33**, 145–168 (2017). doi: 10.1146/annurev-cellbio-100616-060847; pmid: 28693388
7. J. J. Lee et al., Autoproteolysis in hedgehog protein biogenesis. *Science* **266**, 1528–1537 (1994). doi: 10.1126/science.7985023; pmid: 7985023
8. J. A. Porter, K. E. Young, P. A. Beachy, Cholesterol modification of hedgehog signaling proteins in animal development. *Science* **274**, 255–259 (1996). doi: 10.1126/science.7985023; pmid: 7985023
9. K. P. Williams et al., Functional antagonists of sonic hedgehog reveal the importance of the N terminus for activity. *J. Cell Sci.* **112**, 4405–4414 (1999). pmid: 10564658
10. Z. Chamoun et al., Skinny hedgehog, an acyltransferase required for palmitoylation and activity of the hedgehog signal. *Science* **293**, 2080–2084 (2001). doi: 10.1126/science.1064437; pmid: 11486055
11. R. B. Pepinsky et al., Identification of a palmitic acid-modified form of human Sonic hedgehog. *J. Biol. Chem.* **273**,

14037–14045 (1998). doi: 10.1074/jbc.273.22.14037; pmid: 9593755
12. J. D. Kohtz et al., N-terminal fatty-acylation of sonic hedgehog enhances the induction of rodent ventral forebrain neurons. *Development* **128**, 2351–2363 (2001). pmid: 11493554
13. R. J. Dawber, S. Hebbes, B. Herpers, F. Docquier, M. van den Heuvel, Differential range and activity of various forms of the Hedgehog protein. *BMC Dev. Biol.* **5**, 21 (2005). doi: 10.1186/1471-213X-5-21; pmid: 16197551
14. J. A. Goetz, S. Singh, L. M. Suber, F. J. Kull, D. J. Robbins, A highly conserved amino-terminal region of sonic hedgehog is required for the formation of its freely diffusible multimeric form. *J. Biol. Chem.* **281**, 4087–4093 (2006). doi: 10.1074/jbc.M511427200; pmid: 16339763
15. V. Marigo, R. A. Davey, Y. Zuo, J. M. Cunningham, C. J. Tabin, Biochemical evidence that patched is the Hedgehog receptor. *Nature* **384**, 176–179 (1996). doi: 10.1038/384176a0; pmid: 8906794
16. H. Tukachinsky, K. Petrov, M. Watanabe, A. Salic, Mechanism of inhibition of the tumor suppressor Patched by Sonic Hedgehog. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E5866–E5875 (2016). doi: 10.1073/pnas.1606719113; pmid: 27647915
17. N. Denef, D. Neubüser, L. Perez, S. M. Cohen, Hedgehog induces opposite changes in turnover and subcellular localization of patched and smoothened. *Cell* **102**, 521–531 (2000). doi: 10.1016/S0092-8674(00)00056-8; pmid: 10966113
18. J. Taipale, M. K. Cooper, T. Maiti, P. A. Beachy, Patched acts catalytically to suppress the activity of Smoothened. *Nature* **418**, 892–897 (2002). doi: 10.1038/nature00989; pmid: 12192414
19. A. Ruiz i Altaba, P. Sánchez, N. Dahmane, Gli and hedgehog in cancer: Tumours, embryos and stem cells. *Nat. Rev. Cancer* **2**, 361–372 (2002). doi: 10.1038/nrc796; pmid: 12044012
20. C. Metcalfe, F. J. de Sauvage, Hedgehog fights back: Mechanisms of acquired resistance against Smoothened

antagonists. *Cancer Res.* **71**, 5057–5061 (2011). doi: 10.1158/0008-5472.CAN-11-0923; pmid: 21771911

21. T. K. Rimkus, R. L. Carpenter, S. Qasem, M. Chan, H. W. Lo, Targeting the Sonic Hedgehog Signaling Pathway: Review of Smoothened and GLI Inhibitors. *Cancers* **8**, 22 (2016). doi: 10.3390/cancers8020022; pmid: 26891329

22. X. Qi, P. Schmiege, E. Coutavas, J. Wang, X. Li, Structures of human Patched and its complex with native palmitoylated sonic hedgehog. *Nature* **560**, 128–132 (2018). doi: 10.1038/s41586-018-0308-7; pmid: 29995851

23. A. Fleet, J. P. Lee, A. Tamachi, I. Javeed, P. A. Hamel, Activities of the Cytoplasmic Domains of Patched-1 Modulate but Are Not Essential for the Regulation of Canonical Hedgehog Signaling. *J. Biol. Chem.* **291**, 17557–17568 (2016). doi: 10.1074/jbc.M116.731745; pmid: 27325696

24. P. A. Beachy, S. G. Hymowitz, R. A. Lazarus, D. J. Leahy, C. Siebold, Interactions between Hedgehog proteins and their binding partners come into view. *Genes Dev.* **24**, 2001–2012 (2010). doi: 10.1101/gad.1951710; pmid: 20844013

25. N. Fuse et al., Sonic hedgehog protein signals not as a hydrolytic enzyme but as an apparent ligand for patched. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 10992–10999 (1999). doi: 10.1073/pnas.96.20.10992; pmid: 10500113

26. R. B. Pepinsky et al., Mapping sonic hedgehog-receptor interactions by steric interference. *J. Biol. Chem.* **275**, 10995–11001 (2000). doi: 10.1074/jbc.275.15.10995; pmid: 10753901

27. J. S. McLellan et al., The mode of Hedgehog binding to Ihog homologues is not conserved across different phyla. *Nature* **455**, 979–983 (2008). doi: 10.1038/nature07358; pmid: 18794898

28. I. Bosanac et al., The structure of SHH in complex with HHIP reveals a recognition role for the Shh pseudo active site in signaling. *Nat. Struct. Mol. Biol.* **16**, 691–697 (2009). doi: 10.1038/nsmb.1632; pmid: 19561609

29. S. Hurwitz, Homeostatic control of plasma calcium concentration. *Crit. Rev. Biochem. Mol. Biol.* **31**, 41–100 (1996). doi: 10.3109/10409239609110595; pmid: 8744955

30. L. Izzi et al., Boc and Gas1 each form distinct Shh receptor complexes with Ptch1 and are required for Shh-mediated cell proliferation. *Dev. Cell* **20**, 788–801 (2011). doi: 10.1016/j.devcel.2011.04.017; pmid: 21664577

31. X. Gong et al., Structural basis for the recognition of Sonic Hedgehog by human Patched1. *Science* **361**, eaas8935 (2018). doi: 10.3109/10409239609110575; pmid: 8744955

32. E. Krissinel, K. Henrick, Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007). doi: 10.1016/j.jmb.2007.05.022; pmid: 17681537

33. F. Lu et al., Identification of NPC1 as the target of U18666A, an inhibitor of lysosomal cholesterol export and Ebola infection. *eLife* **4**, e12177 (2015). doi: 10.7554/eLife.12177; pmid: 26646182

34. X. Li et al., Structure of human Niemann-Pick C1 protein. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 8212–8217 (2016). doi: 10.1073/pnas.1607795113; pmid: 27307437

35. X. Li, P. Saha, J. Li, G. Blobel, S. R. Pfeffer, Clues to the mechanism of cholesterol transfer from the structure of NPC1 middle lumenal domain bound to NPC2. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 10079–10084 (2016). doi: 10.1073/pnas.1611956113; pmid: 27551080

36. B. L. Allen et al., Overlapping roles and collective requirement for the coreceptors GAS1, CDO, and BOC in SHH pathway function. *Dev. Cell* **20**, 775–787 (2011). doi: 10.1016/j.devcel.2011.04.018; pmid: 21664576

37. J. Ericson, S. Morton, A. Kawakami, H. Roelink, T. M. Jessell, Two critical periods of Sonic Hedgehog signaling required for the specification of motor neuron identity. *Cell* **87**,

661–673 (1996). doi: 10.1016/S0092-8674(00)81386-0; pmid: 8929535

38. H. R. Maun et al., Hedgehog pathway antagonist 5E1 binds hedgehog at the pseudo-active site. *J. Biol. Chem.* **285**, 26570–26580 (2010). doi: 10.1074/jbc.M110.112284; pmid: 20504762

39. P. T. Chuang, T. Kawcak, A. P. McMahon, Feedback control of mammalian Hedgehog signaling by the Hedgehog-binding protein, Hip1, modulates Fgf signaling during branching morphogenesis of the lung. *Genes Dev.* **17**, 342–347 (2003). doi: 10.1101/gad.1026303; pmid: 12569124

40. B. Bishop et al., Structural insights into hedgehog ligand sequestration by the human hedgehog-interacting protein HHIP. *Nat. Struct. Mol. Biol.* **16**, 698–703 (2009). doi: 10.1038/nsmb.1607; pmid: 19561611

41. K. Berka et al., MOLEonline 2.0: Interactive web-based analysis of biomacromolecular channels. *Nucleic Acids Res.* **40**, W222–W227 (2012). doi: 10.1093/nar/gks363; pmid: 22553366

42. M. H. Chen, Y. J. Li, T. Kawakami, S. M. Xu, P. T. Chuang, Palmitoylation is required for the production of a soluble multimeric Hedgehog protein complex and long-range signaling in vertebrates. *Genes Dev.* **18**, 641–659 (2004). doi: 10.1101/gad.1185804; pmid: 15075292

43. K. Amanai, J. Jiang, Distinct roles of Central missing and Dispatched in sending the Hedgehog signal. *Development* **128**, 5119–5127 (2001). pmid: 11748147

44. J. D. Lee, J. E. Treisman, Sightless has homology to transmembrane acyltransferases and is required to generate active Hedgehog protein. *Curr. Biol.* **11**, 1147–1152 (2001). doi: 10.1016/S0960-9822(01)00323-2; pmid: 11509241

45. D. J. Stauber, E. W. Debler, P. A. Horton, K. A. Smith, I. A. Wilson, Crystal structure of the IL-2 signaling complex: Paradigm for a heterotrimeric cytokine receptor. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 2788–2793 (2006). doi: 10.1073/pnas.0511161103; pmid: 16477002

46. Y. Bellaiche, I. The, N. Perrimon, Tout-velu is a Drosophila homologue of the putative tumour suppressor EXT-1 and is needed for Hh diffusion. *Nature* **394**, 85–88 (1998). doi: 10.1038/27932; pmid: 9665133

47. I. The, Y. Bellaiche, N. Perrimon, Hedgehog movement is regulated through tout velu-dependent synthesis of a heparan sulfate proteoglycan. *Mol. Cell* **4**, 633–639 (1999). doi: 10.1016/S1097-2765(00)80214-2; pmid: 10549295

48. B. L. Allen, T. Tenzen, A. P. McMahon, The Hedgehog-binding proteins Gas1 and Cdo cooperate to positively regulate Shh signaling during mouse development. *Genes Dev.* **21**, 1244–1257 (2007). doi: 10.1101/gad.1543607; pmid: 17504941

49. S. Yao, L. Lum, P. Beachy, The ihog cell-surface proteins bind Hedgehog and mediate pathway activation. *Cell* **125**, 343–357 (2006). doi: 10.1016/j.cell.2006.02.040; pmid: 16630821

50. J. Y. Song, A. M. Holtz, J. M. Pinskey, B. L. Allen, Distinct structural requirements for CDON and BOC in the promotion of Hedgehog signaling. *Dev. Biol.* **402**, 239–252 (2015). doi: 10.1016/j.ydbio.2015.03.015; pmid: 25848697

51. X. Li et al., Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat. Methods* **10**, 584–590 (2013). doi: 10.1038/nmeth.2472; pmid: 23644547

52. A. Rohou, N. Grigorieff, CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015). doi: 10.1016/j.jsb.2015.08.008; pmid: 26278980

53. S. H. Scheres, RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012). doi: 10.1016/j.jsb.2012.09.006; pmid: 23000701

54. N. Grigorieff, Frealign: An Exploratory Tool for Single-Particle Cryo-EM. *Methods Enzymol.* **579**, 191–226 (2016). doi: 10.1016/bs.mie.2016.04.013; pmid: 27572728

55. P. D. Adams et al., PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010). doi: 10.1107/S0907444909052925; pmid: 20124702

56. G. N. Murshudov, A. A. Vagin, E. J. Dodson, Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997). doi: 10.1107/S0907444996012255; pmid: 15299926

57. A. Brown et al., Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Crystallogr. D* **71**, 136–153 (2015). doi: 10.1107/S1399004714021683; pmid: 25615868

58. L. F. Ten Eyck, Efficient structure-factor calculation for large molecules by the fast fourier transform. *Acta Crystallogr. A* **33**, 486–492 (1977). doi: 10.1107/S0567739477001211

59. Z. Wang et al., An atomic model of brome mosaic virus using direct electron detection and real-space optimization. *Nat. Commun.* **5**, 4808 (2014). doi: 10.1038/ncomms5808; pmid: 25185801

60. J. B. Heymann, D. M. Belnap, Bsoft: Image processing and molecular modeling for electron microscopy. *J. Struct. Biol.* **157**, 3–18 (2007). doi: 10.1016/j.jsb.2006.06.006; pmid: 17011211

61. V. B. Chen et al., MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010). doi: 10.1107/S0907444909042073; pmid: 20057044

62. E. F. Pettersen et al., UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004). doi: 10.1002/jcc.20084; pmid: 15264254

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/362/6410/eaas8843/suppl/DC1
Figs. S1 to S5
Table S1

REPORT

Glacial lake outburst floods as drivers of fluvial erosion in the Himalaya

Kristen L. Cook[1]*, Christoff Andermann[1], Florent Gimbert[1,2],
Basanta Raj Adhikari[3], Niels Hovius[1,4]

Himalayan rivers are frequently hit by catastrophic floods that are caused by the failure of glacial lake and landslide dams; however, the dynamics and long-term impacts of such floods remain poorly understood. We present a comprehensive set of observations that capture the July 2016 glacial lake outburst flood (GLOF) in the Bhotekoshi/Sunkoshi River of Nepal. Seismic records of the flood provide new insights into GLOF mechanics and their ability to mobilize large boulders that otherwise prevent channel erosion. Because of this boulder mobilization, GLOF impacts far exceed those of the annual summer monsoon, and GLOFs may dominate fluvial erosion and channel-hillslope coupling many tens of kilometers downstream of glaciated areas. Long-term valley evolution in these regions may therefore be driven by GLOF frequency and magnitude, rather than by precipitation.

Lake outburst floods (LOFs) have long been recognized as both a hazard and major agent of geomorphic change in the Himalaya (1–5). These floods originate from lakes that have formed behind a landslide dam or in association with a glacier, dammed by a frontal moraine or glacial ice. Such lakes can drain catastrophically for several reasons, including mass movements or avalanches into the lake, seismic activity, piping within the dam, overtopping of the dam, or degradation of blocking ice (4, 6). The resulting floods can have short-lived discharges up to several orders of magnitude higher than background discharges in the receiving rivers (7). Because of their magnitude and unpredictability, LOFs can be highly destructive and compromise local infrastructure such as roads, buildings, and hydropower facilities (3, 8–10).

Although large LOFs have been recognized as strongly affecting river morphology and dynamics (4, 11–13), they are often treated as one-off events. The potential impact of repeated LOFs, particularly the less dramatic small-to-medium magnitude floods, on the longer-term behavior of the fluvial system has received little attention. The impact of individual LOF events must be considered along with LOF frequency and measured against the accumulated effect of annual monsoon floods of variable size. We evaluate the importance of glacial lake outburst floods (GLOFs) in driving fluvial erosion by examining the Bhotekoshi/Sunkoshi River, where we compare monsoon floods to a GLOF that occurred in July 2016. In addition to documenting relative impacts and discharges, we use seismic observations to gain insight into GLOF dynamics, and we explore the role of boulder-sized sediment in promoting GLOF-driven erosion.

On the night of 5 July 2016, Gongbatongshacuo Lake, a 1.7×10^4-m^2 moraine-dammed lake in the Tibet Autonomous Region, China, drained catastrophically, releasing approximately 1.1×10^5 m^3 of water (14). The cause of the breach is unknown, but fresh deposits above the lake suggest that it may have been associated with a debris flow event, possibly increasing the volume of the flood (Fig. 1B). The flood proceeded down the Zhangzangbo River into the Poiqu/Bhotekoshi/Sunkoshi River and caused severe damage in the Bhotekoshi valley, destroying the intake dam of a hydropower project, the Araniko highway, and numerous buildings in the towns of Kodari and Tatopani. The zone of damage sits within the area affected by strong ground motion and landsliding induced by the 2015 moment magnitude 7.8 Gorkha earthquake, which had an estimated return time of a few hundred years (15, 16).

The 2016 GLOF passed through an array of six broadband seismometers installed in 2015 along the Bhotekoshi River valley, 28 to 35 km downstream of the Zhangzangbo confluence (Fig. 1). Because both turbulent flow (17) and bedload transport (18) in rivers can generate detectable seismic ground motion (19), the seismic record of the GLOF can be used to probe the flood and sediment dynamics at high temporal and spatial resolution. The seismic records for each near-river station contain two distinct pulses of high-amplitude noise (Fig. 1C). The first is the flood front, which propagates between stations at 8.7 m/s. The abrupt rise of seismic power at each station suggests that the maximum flow depth was reached within 2 min. At Chaku station, this corresponded to a calculated maximum discharge of 1500 to 2100 m^3/s (Fig. 2 and table S1). The second pulse is of higher magnitude but travels slower, at 5 m/s. The total duration of the flood, including both pulses, at each of the stations was less than an hour. We interpret the first pulse as a water wave and the second as a package of coarse sediment, on the basis of the following reasoning: If both pulses were associated with propagating water waves, then the more energetic second wave should have a greater flow depth and travel faster (17). The prediction from theory is that near-river stations should be especially sensitive to coarse sediment, whereas far-river stations should be predominantly sensitive to water flow (17, 18) (fig. S1). The second wave was prominent at near-river stations but much weaker at farther seismic stations, consistent with the theory (fig. S2). Theory also predicts that coarse-sediment transport generates seismic noise at higher frequencies compared to turbulent flow (17, 18) (fig. S1), and power spectra from the two pulses indicate an increase in power at higher frequencies during the second pulse (fig. S3).

The ratio of the pulse velocities is 0.6, matching bedload/water velocity ratios observed experimentally and in small-river settings (20). This velocity difference ensured that the flood front outpaced any entrained sediment and therefore remained depleted in bedload.

Field- and satellite-based observations show that the GLOF affected the river channel over a ~40-km stretch between the confluence with the Zhangzangbo River and Barabise town (Fig. 1). The flood impact extended into the adjacent hillslopes through undercutting and destabilization of the river banks, leading to bank collapses, slumps, and landslides. The extensive flood-induced damage to local infrastructure was almost exclusively the result of bank erosion and mass wasting, rather than inundation (fig. S4).

We quantified the magnitude of channel, bank, and hillslope change with repeat terrestrial lidar surveys from October 2015, March 2016, and November 2016 in nine locations that together covered 20% of the channel length between Khukundol and Barabise (Fig. 1 and table S2). Eight of the scanned reaches experienced downcutting varying from 1 to 10 m, whereas one had no bed-elevation change (Fig. 3).

We observed bank erosion in numerous sections of the channel, and all scanned reaches contained segments with at least 2 m of lateral erosion (Fig. 3 and fig. S5). Bank erosion included parallel retreat of steep banks and erosion and undercutting at the base of slopes. Analysis of 5-m-resolution RapidEye imagery (table S3) indicates that the 2016 GLOF caused the mean width of the active channel between the Zhangzangbo confluence and Barabise to increase from 29.5 ± 3 m

[1]GFZ German Research Centre for Geosciences, Potsdam, Germany. [2]University of Grenoble Alpes, CNRS, IRD, Institut des Géosciences de l'Environnement (IGE), Grenoble, France. [3]Department of Civil Engineering, Pulchowk Campus, Institute of Engineering, Tribhuvan University, Nepal. [4]Institute of Earth and Environmental Science, Potsdam University, Potsdam, Germany.
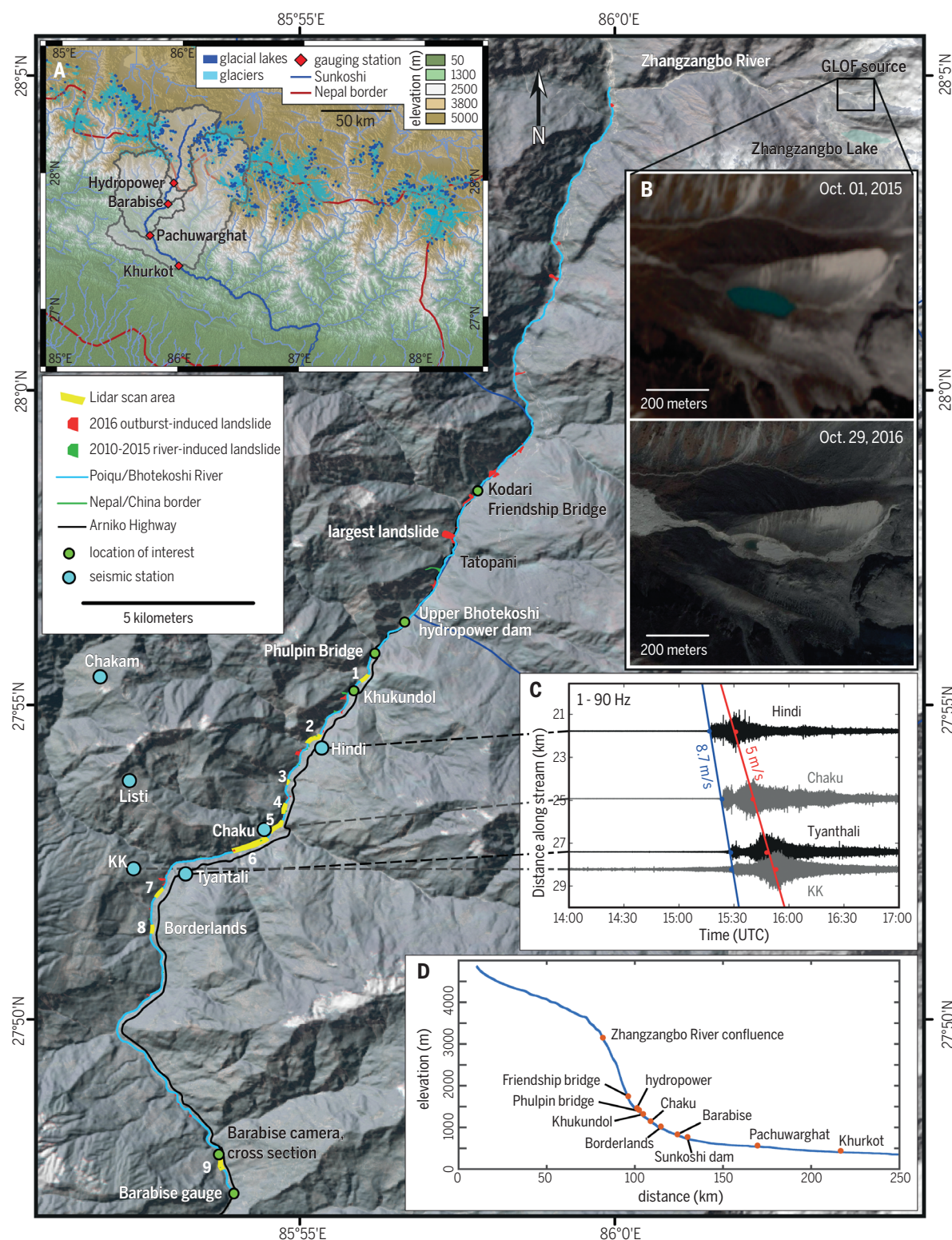*Corresponding author. Email: klcook@gfz-potsdam.de

Fig. 1. Map of the Bhotekoshi study region. Scanned reaches, mapped landslides, seismic stations, and locations discussed in the text are shown. White numbers 1 to 9 indicate each scan location, with reference to table S1. (**A**) The inset provides the regional context, with glaciers and glacial lakes shown (*9*). The shading and black outlines indicate drainage basins upstream of Khurkot, Pachuwarghat, Barabise, and the Upper Bhotekoshi hydropower dam. (**B**) Google Earth and RapidEYE imagery showing a magnified view of the lake that was the source of the outburst flood, both before the bursting event in October 2015 (top) and after in October 2016 (bottom). (**C**) Seismic record of GLOF propagation. Normalized ground velocity time series from four stations at different distances downstream of the Zhangzangbo confluence. Dots indicate the manually picked pulse arrivals. Straight lines correspond to linear fits of distance versus time and yield the pulse velocities. UTC, universal time coordinated. (**D**) Longitudinal profile from the Advanced Land Observing Satellite 12.5-m digital elevation model of the Poiqu/Bhotekoshi/Sunkoshi River, with locations of interest marked.

in 2015 to 41.3 ± 3 m in 2016, with highly variable widening throughout the mapped area (Fig. 3). The lateral changes in 2016 contrast with the stability of the river between 2010 and 2015. During these six monsoon seasons, changes to the river channel were minimal and were associated with external influences such as anthropogenic modification, local landsliding, and tributary input. Despite the large amount of landslide debris produced during the 2015 Gorkha earthquake (15, 16), the channel underwent minimal modification during the 2015 monsoon (Fig. 3 and fig. S6).

Lateral erosion of the channel by the GLOF led to the activation of landslides that propagated
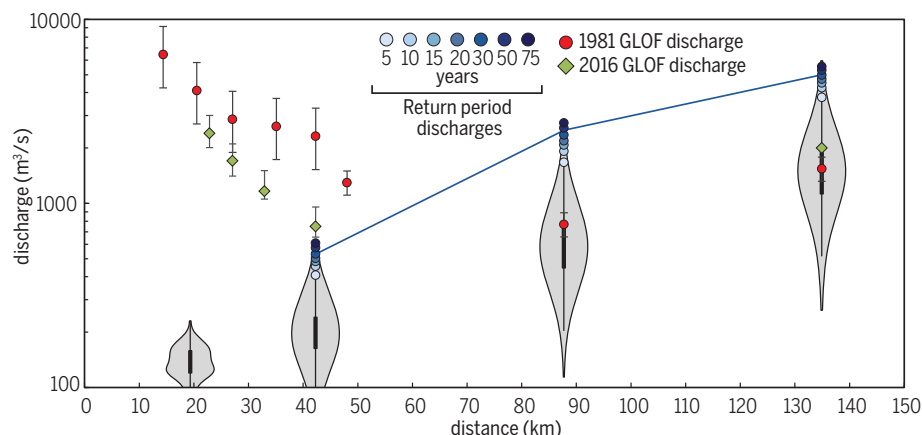
Fig. 2. Discharge with distance downstream of the Zhangzangbo confluence. Data from table S4. Estimated discharges for the 1981 (11) and 2016 GLOFs and the calculated discharges for floods of varying return periods (5 to 75 years). The blue line connects the 30-year-return discharges. The violin plots show the full distribution of monsoon discharges (July and August) for each hydrological station (hydropower dam, Barabise, Pachuwarghat, and Khurkot). Error bars indicate estimated uncertainty (for details, see materials and methods).
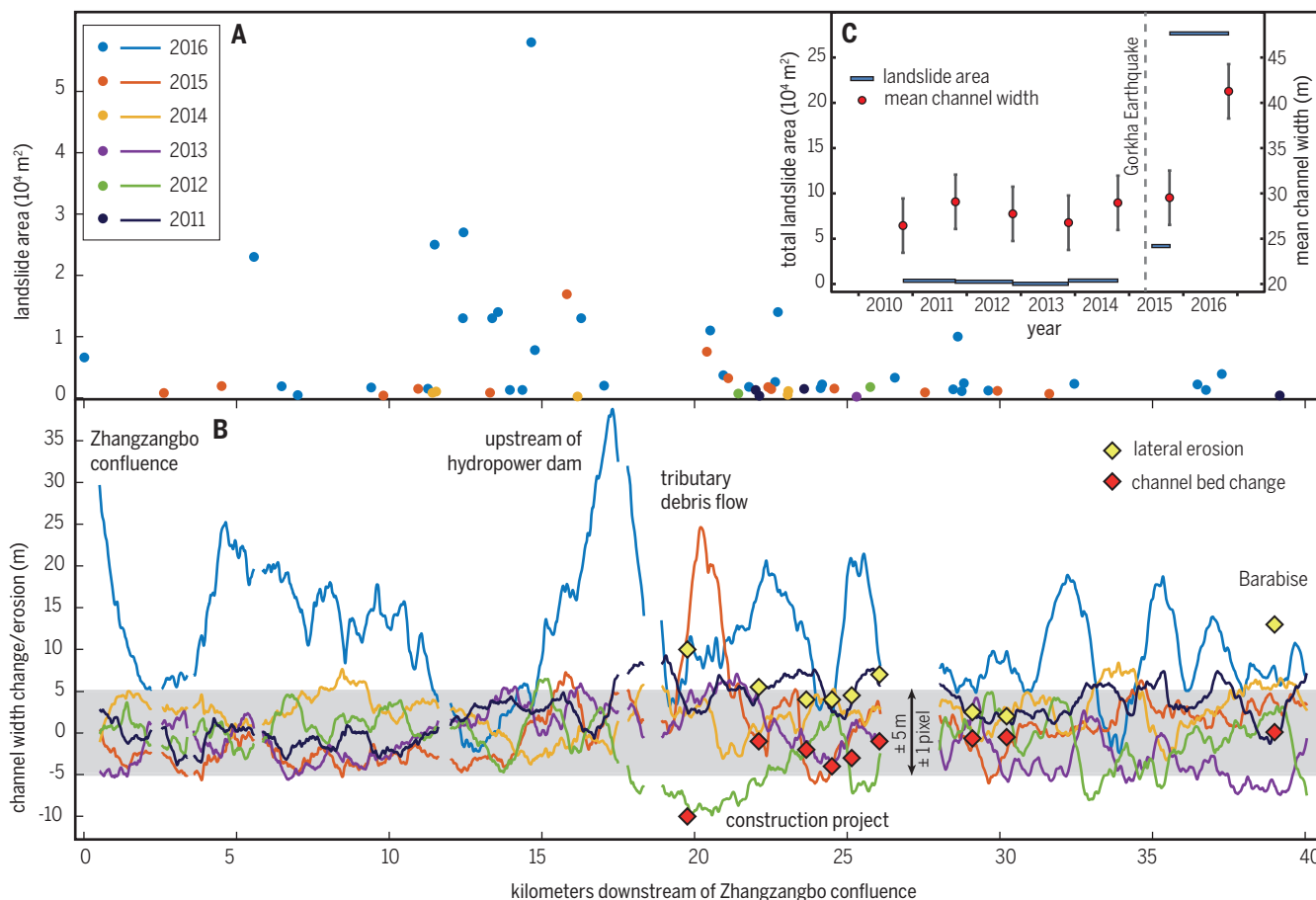
Fig. 3. Summary of monsoon- and GLOF-driven changes in the Poiqu/Bhotekoshi channel. (**A**) Landslide areas with distance downstream of the Zhangzangbo confluence from 2010 to 2016; each point represents one landslide. (**B**) Width changes from 2010 to 2016 and lidar-derived GLOF changes. The gray bar indicates an uncertainty of ±5 m, equivalent to ±1 pixel in the satellite imagery. Data have been smoothed by applying a running average over a 1-km window. Diamonds show the maximum values of lateral erosion and channel-bed elevation change from each of the lidar-scanned reaches. The lidar-derived lateral erosion values are not expected to match the satellite-derived width changes because the satellite-derived data have been smoothed. (**C**) Time series summary of data from (A) and (B). The mean channel widths through time and areas of river-related landslides that occurred during each monsoon season are shown.

up the hillslopes (Fig. 1). Mapping from RapidEye imagery shows that 26 landslides in which the zone of failure is connected to the channel formed during the 2016 monsoon. The cumulative area of these landslides was ~100 times larger than the landslide area during a typical year (2009 to 2014) and five times larger than that during the 2015 monsoon, which solicited an unusually high rate of landsliding after the Gorkha earthquake (Figs. 1 and 3).

The 2016 GLOF had an impact on the river and adjacent hillslopes that far outstripped that of the monsoon floods of 2009 to 2016. The effects of the 2016 GLOF are similar to the documented effects of a previous GLOF on the Poiqu/Bhotekoshi/Sunkoshi River on 11 July 1981. This flood contained an estimated 1.9×10^7 m^3 of water released from the Zhangzangbo Lake, near the source of the 2016 flood (11) (Fig. 1). The damage from the 1981 flood mirrors that from the 2016 flood and included the destruction of a hydropower dam, sections of the Araniko highway, and several bridges. Numerous landslides related to this outburst flood occurred along the Bhotekoshi River (11).

We attribute the large discrepancy between the impacts of the GLOFs and the monsoon floods to the ubiquity of extremely coarse landslide-derived boulders (>1 m in diameter) in the bed and banks of the Bhotekoshi River (fig. S7). River-channel stability during floods is related to the stability of boulder-sized clasts that define the channel geometry (21, 22). Although the threshold discharge for mobilizing these boulders varies with clast location and size, the threshold generally does not appear to be exceeded during the monsoon. On the basis of field observations and Google Earth imagery, movement of large boulders has not taken place during monsoon floods since at least 2004 (fig. S8). The smallest boulders we can reliably identify as stable in Google Earth imagery are about 2 to 3 m in diameter, smaller than boulders that moved in these reaches during the 2016 and 1981 GLOFs [up to 5.7 m (fig. S5) and 13.4 m in diameter, respectively (11)].

GLOFs can mobilize boulders, owing to their high discharge and other characteristics that enhance sediment entrainment. Outburst floods, in which a water bore propagates downstream, have a higher capacity to mobilize sediment than a monsoon flood of similar magnitude. This is due to the velocity difference between water flow and entrained bedload, which ensures that the leading edge of the flood will remain relatively bedload free and under transport capacity. This is fundamentally different from run-off–driven floods, which have more smoothly varying hydrographs and sediment loads delivered from outside the channel.

The ability to mobilize the channel-defining coarse sediment determines the degrees to which a flood can incise bedrock and erode the channel banks. The bedrock bed of the observed section of the Bhotekoshi River is covered by a sediment layer of unknown thickness. A flood that does not move the large boulders armoring this sediment mantle therefore cannot cause bedrock incision.
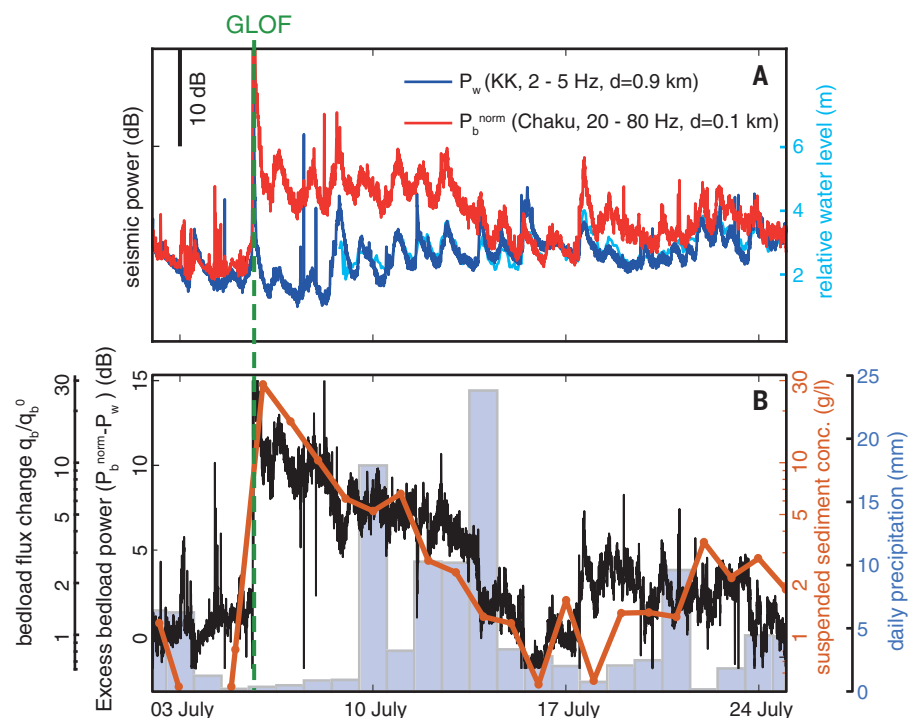
Fig. 4. Sediment dynamics after the GLOF. (**A**) Constraints on bedload transport. The dark blue line shows the seismic power in the 2- to 5-Hz frequency range (P_w) at station KK, a distance (d) of 0.9 km from the river, as a proxy for turbulent flow. The red line shows the seismic power in the 20- to 80-Hz frequency range (P_b^{norm}) at station Chaku, 0.1 km from the river, as a proxy for bedload flux. The light blue line shows the water level relative to the Nepal Department of Hydrology and Meteorology gauge at Barabise. Bedload power has been normalized to the turbulent flow power by using the period before the GLOF. Note that seismic power during the GLOF event itself is off the scale of this plot. (**B**) Excess sediment transport. The black line shows the excess bedload power, obtained by differencing the blue and red series in (A). The GLOF event is off the scale of this plot and is not considered in this analysis. The orange line shows the suspended-sediment concentration from daily samples. The bars show Global Precipitation Monitoring–derived catchment-wide daily precipitation. q_b, bedload flux; q_b^0, background bedload flux.

The impact of the 2016 and 1981 GLOFs on the bedrock below the sediment layer cannot be constrained; however, we can conclude that these GLOFs may have incised bedrock, whereas the monsoon floods since at least 2004 definitely did not.

GLOF-induced disruption of the boulder armor is reflected in increased rates of sediment transport during the 10 days after the 2016 GLOF. We used seismic signals in different frequency ranges from a far-river station and a near-river station to obtain proxies for water flow depth and bedload flux (17, 18) (Fig. 4A and fig. S1). Before the GLOF, the bedload flux and water depth proxies are closely linked. After the GLOF, however, the two proxies show different trends. Whereas the flow depth proxy returns to preflood levels within hours, the bedload transport proxy is increased after the flood and gradually returns to preflood levels over 10 days (Fig. 4B). Daily suspended-sediment concentrations from Barabise show a very similar perturbation, with changes of the same magnitude and time evolution as the estimated excess bedload transport (Fig. 4B), suggest-

ing that both fluxes are controlled by sediment availability from the same source. We interpret this as a signal of ongoing reorganization of the channel bed; the return to background levels of sediment transport occurs as easily transportable sediment is removed from the channel and the boulder armor reestablishes. Furthermore, the increase in bedload transport after the GLOF suggests that at most other times the river is under capacity and that bedload transport during the monsoon is typically limited by the delivery of sediment into the channel.

Because of the high discharge threshold for mobilizing the channel-defining boulders, monsoon floods primarily transfer sediment delivered from tributaries and hillslopes without major lateral or vertical erosion in the main channel, and only very large floods cross the threshold for boulder mobilization, perturb the river, and drive erosion. This allows individual floods, including GLOFs and landslide dam outbursts, to have a disproportionate impact on the river channel. The role of GLOFs in driving long-term erosion and hillslope-channel coupling therefore depends

on their frequency and magnitude relative to that of extreme monsoon-driven floods.

The frequency of GLOFs in the central Himalaya is difficult to establish, because records are incomplete and recorded floods may not be correctly identified as GLOFs (*3, 23*). Nevertheless, GLOFs are relatively common in the Himalaya, with a major flood occurring at least once every 2 years on average (*4, 24–26*). The Bhotekoshi River has experienced GLOFs in 1935, 1964, 1981, and 2016, suggesting a return period of about 30 years (*2*). The Bhotekoshi/Sunkoshi catchment has 57 glacial lakes mapped upstream of Barabise (*10*). These lakes vary widely in size, but the median lake area of 32,600 m^2 is almost twice that of the lake that caused the 2016 flood. Yearly maximum monsoon discharge from a 42-year discharge record at Barabise is typically between 200 and 400 m^3/s, with flood peaks rarely exceeding 500 m^3/s (fig. S9). The estimated GLOF peak discharges at Barabise [700 to 900 m^3/s in 2016 and 2300 m^3/s in 1981 (*11*)] are larger than the expected 30-year flood discharge of 490 to 560 m^3/s (Fig. 2, fig. S11, and table S4). The impact of individual GLOFs on the Bhotekoshi River dwarfs that of monsoon discharges, and GLOFs occur with sufficient frequency to dominate geomorphic change in the valley.

As a GLOF travels downstream, the flood peak attenuates and the peak discharge decreases, whereas the drainage area and background discharge increase, giving rise to a crossover point where GLOF discharges are no longer anomalous (*7, 10*) (Fig. 2). Hence, the discrepancy between GLOF discharge and monsoon floods will be particularly dramatic in the high Himalayan headwaters, where the drainage areas are small and GLOFs will have experienced little attenuation (Fig. 2). For the 2016 and 1981 GLOFs, the crossover point with the 30-year monsoon flood was located about 45 and 55 km downstream of the Zhangzangbo confluence, respectively (Fig. 1). If GLOFs are less frequent, then their impact must be measured against larger monsoon floods with a longer return time, moving the crossover point upstream. Conversely, larger GLOFs have a crossover point farther downstream. Probabilistic modeling of glacial lake outbursts throughout the Himalaya suggests that more than 40% of 2359 mapped proglacial lakes could produce GLOFs that match the 100-year discharge about 20 km downstream, whereas large GLOFs may reach as far as 85 km downstream (*10*). Our observations suggest that, because of their distinct sediment dynamics, GLOFs of such magnitudes will have a disproportionate effect on fluvial erosion in these reaches.

Owing to their magnitude and enhanced ability to mobilize coarse sediment, we propose that LOFs are a fundamental part of the fluvial system and a primary control on fluvial erosion and channel-hillslope coupling, especially in catchments where very coarse sediment creates high thresholds for sediment mobilization, with GLOFs particularly effective in the upper portions of glaciated catchments. Landslide LOFs likely have a similar impact on channel dynamics and channel-hillslope coupling in landslide-prone regions with steep narrow valleys and abundant coarse sediment, conditions that are common in numerous mountain ranges throughout the world (*4, 5*).

LOFs directly impact only the channel and adjacent hillslopes, but over the long term, fluvial incision sets the base level for the entire landscape and is ultimately the driver of hillslope erosion; thus, in LOF-prone regions, LOF magnitudes and recurrence intervals may control landscape evolution. As a result, monsoon strength or measures of precipitation may be poor predictors of landscape response in LOF-susceptible regions. Instead, erosion rates may be strongly influenced by nonclimatological LOF drivers such as earthquakes (*5, 27, 28*) and the climatic factors that affect the size and distribution of glacial lakes, for example, air temperature, variability of the equilibrium line altitude, and, to a degree, the glacial recharge (*29, 30*). Even where LOF frequency can be linked to precipitation (that is, for landslide lake outbursts), the relationship between fluvial erosion and precipitation will become complicated and nonlinear.

A warming climate is thought to promote glacial lake formation in some areas as retreating glaciers create space for lakes behind abandoned end moraines and increased melting rates supply more water to potential lakes. This, in turn, may increase GLOF frequency and/or magnitude (*25, 31–33*). The potential for increased GLOF activity in response to climate change therefore not only represents increased risk to communities in these regions but may also strongly affect the pace of landscape change in a way that is not reflected in precipitation-dependent erosion models.

REFERENCES AND NOTES

1. K. Hewitt, "Natural dams and outburst floods of the Karakoram Himalaya," in *Hydrological Aspects of Alpine and High-Mountain Areas*, J. W. Glen, Ed. (IAHS Publication No. 138, International Association of Hydrological Sciences, 1982), pp. 259–269.
2. P. K. Mool, *J. Nepal Geol. Soc.* **11**, 273–280 (1995).
3. International Centre for Integrated Mountain Development (ICIMOD), "Glacial lakes and glacial lake outburst floods in Nepal" (ICIMOD, Kathmandu, 2011).
4. O. Korup, F. Tweed, *Quat. Sci. Rev.* **26**, 3406–3422 (2007).
5. J. E. Costa, R. L. Schuster, *Geol. Soc. Am. Bull.* **100**, 1054–1068 (1988).
6. M. J. Westoby et al., *Earth Sci. Rev.* **134**, 137–159 (2014).
7. D. A. Cenderelli, E. E. Wohl, *Geomorphology* **40**, 57–90 (2001).
8. A. B. Shrestha et al., *Geomatics Nat. Hazards Risk* **1**, 157–169 (2010).
9. N. R. Khanal, J. M. Hu, P. Mool, *Mt. Res. Dev.* **35**, 351–364 (2015).
10. W. Schwanghart, R. Worni, C. Huggel, M. Stoffel, O. Korup, *Environ. Res. Lett.* **11**, 074005 (2016).
11. D. Xu, *GeoJournal* **17**, 569–580 (1988).
12. D. A. Cenderelli, E. E. Wohl, *Earth Surf. Process. Landf.* **28**, 385–407 (2003).
13. D. Higaki, G. Sato, *Glob. Environ. Res.* **16**, 71–76 (2012).
14. C. Huggel, A. Kääb, W. Haeberli, P. Teysseire, F. Paul, *Can. Geotech. J.* **39**, 316–330 (2002).
15. J. S. Kargel et al., *Science* **351**, aac8353 (2016).
16. K. Roback et al., *Geomorphology* **301**, 121–138 (2017).
17. F. Gimbert, V. C. Tsai, M. P. Lamb, *J. Geophys. Res. Earth Surf.* **119**, 2209–2238 (2014).
18. V. C. Tsai, B. Minchew, M. P. Lamb, J. P. Ampuero, *Geophys. Res. Lett.* **39**, L02404 (2012).
19. A. Burtin, L. Bollinger, J. Vergne, R. Cattin, J. L. Nábělek, *J. Geophys. Res. Solid Earth* **113**, B05301 (2008).
20. P. Chatanantavet, K. X. Whipple, M. A. Adams, M. P. Lamb, *J. Geophys. Res. Earth Surf.* **118**, 1161–1176 (2013).
21. D. A. Cenderelli, B. L. Cluer, "Depositional processes and sediment supply in resistant-boundary channels: Examples from two case studies," in *Rivers Over Rock: Fluvial Processes in Bedrock Channels*, K. J. Tinkler, E. E. Wohl, Eds. (Geophysical Monograph Series, vol. 107, American Geophysical Union, Washington, DC, 1998), pp. 105–131.
22. A. C. Brayshaw, *Geol. Soc. Am. Bull.* **96**, 218–223 (1985).
23. Rastriya Samachar Samiti, "Heavy rain brings massive flood to Bhotekoshi River," *Himalyan Times*, 6 July 2016; https://thehimalayantimes.com/nepal/heavy-rain-brings-massive-flood-bhotekoshi-river.
24. J. D. Ives, R. B. Shrestha, P. K. Mool, "Formation of glacial lakes in the Hindu Kush-Himalayas and GLOF risk assessment" (ICIMOD, Kathmandu, 2010).
25. A. Lutz, W. W. Immerzeel, S. R. Bajracharya, M. Litt, A. Shrestha, "Impact of climate change on the cryosphere, hydrological regimes and glacial lakes of the Hindu Kush Himalayas: A review of current knowledge" (ICIMOD Research Report 2016/3, ICIMOD, Kathmandu, 2016).
26. D. R. Rounce, A. C. Byers, E. A. Byers, D. C. McKinney, *Cryosphere* **11**, 443–449 (2017).
27. W. Schwanghart et al., *Science* **351**, 147–150 (2016).
28. J. D. Bricker et al., *Mt. Res. Dev.* **37**, 5–15 (2017).
29. S. G. Evans, J. J. Clague, *Geomorphology* **10**, 107–128 (1994).
30. K. Hewitt, J. Liu, *Phys. Geogr.* **31**, 528–551 (2013).
31. S. R. Bajracharya, P. Mool, *Ann. Glaciol.* **50**, 81–86 (2009).
32. T. Bolch et al., *Science* **336**, 310–314 (2012).
33. Y. Chen, C. Xu, Y. Chen, W. Li, J. Liu, *Quat. Int.* **226**, 75–81 (2010).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/362/6410/53/suppl/DC1
Materials and Methods
Figs. S1 to S12
Tables S1 to S6
References (*34–38*)

6 March 2018; accepted 3 August 2018
10.1126/science.aat4981

GEOPHYSICS

Slab2, a comprehensive subduction zone geometry model

Gavin P. Hayes[1]*, Ginevra L. Moore[1,2]†, Daniel E. Portner[1,3], Mike Hearne[1],
Hanna Flamme[1,2]‡, Maria Furtney[1]§, Gregory M. Smoczyk[1]

Subduction zones are home to the most seismically active faults on the planet. The
shallow megathrust interfaces of subduction zones host Earth's largest earthquakes
and are likely the only faults capable of magnitude 9+ ruptures. Despite these facts, our
knowledge of subduction zone geometry—which likely plays a key role in determining
the spatial extent and ultimately the size of subduction zone earthquakes—is incomplete.
We calculated the three-dimensional geometries of all seismically active global
subduction zones. The resulting model, called Slab2, provides a uniform geometrical
analysis of all currently subducting slabs.

Subduction zones account for about ~90% of
the energy released by historical earthquakes
globally. The world's largest earthquakes
occur there, and often such events are both
damaging and deadly. Knowledge of the
geometry of subducting slabs is important to
the understanding of earthquake and tsunami
hazards associated with subduction zones. Slab1.0
(*1*), the predecessor to the model presented here,
was introduced as a three-dimensional (3D) model
of the geometries of subduction zones worldwide
that synthesized results from global earthquake
monitoring and regional seismotectonic studies

and provided data at a resolution useful for seis-
mic and tsunami hazard studies. Many authors
have used Slab1.0 as a reference for earthquake
source characterization (*2*, *3*). The geometries
provided by the model have been important for
global earthquake hazard analysis (*4*). The model
has also been used in studies of geophysical pro-
cesses, such as megathrust behavior (*5*, *6*), slow
slip (*7*), seismic anisotropy (*8*), volcano petrology
(*9*), tsunami hazard (*10*), and mantle flow (*11*).
However, some aspects of Slab1.0 are now out of
date with respect to current knowledge of sub-
duction zone structure. Moreover, the approaches

used in Slab1.0 proved inadequate for some im-
portant subduction zones, such as those in the
Himalaya, eastern Indonesia, Manila, and New
Zealand. Though several other global subduction
zone geometry databases exist, most are lim-
ited in either spatial extent or resolution (*12*, *13*)
(fig. S1), particularly at shallow depths (<50 km),
where the largest, most hazardous subduction
zone earthquakes ("megathrust" events) occur
and where most tsunamigenic earthquakes are
located. As tsunamis are very sensitive to the
geometry of the fault on which slip occurred, de-
tailed knowledge of the shape of the megathrust
is critical for hazard assessment.

Here, we present Slab2, a new model describ-
ing the 3D geometries of all seismically active
subduction zones worldwide (Fig. 1). We model
slabs from the near-surface (oceanic trenches for
most slabs) to their deepest expressions in the
upper mantle. In all, Slab2 describes the detailed
geometry of more than 24 million square kilome-
ters of subducted slabs. We made a variety of
enhancements to Slab2 that set it apart from
Slab1.0 and other global subduction zone models.
First and foremost, we collected more data that
image subducting slabs. Data sources come from

[1]U.S. Geological Survey National Earthquake Information
Center, Golden, CO, USA. [2]Department of Geophysics,
Colorado School of Mines, Golden, CO, USA. [3]Department of
Geosciences, University of Arizona, Tucson, AZ, USA.
*Corresponding author. Email: ghayes@usgs.gov †Present
address: School of Oceanography, University of Washington,
Seattle, WA, USA. ‡Present address: Chevron, Houston, TX, USA.
§Present address: Department of Earth, Environmental and
Planetary Sciences, Rice University, Houston, TX, USA.
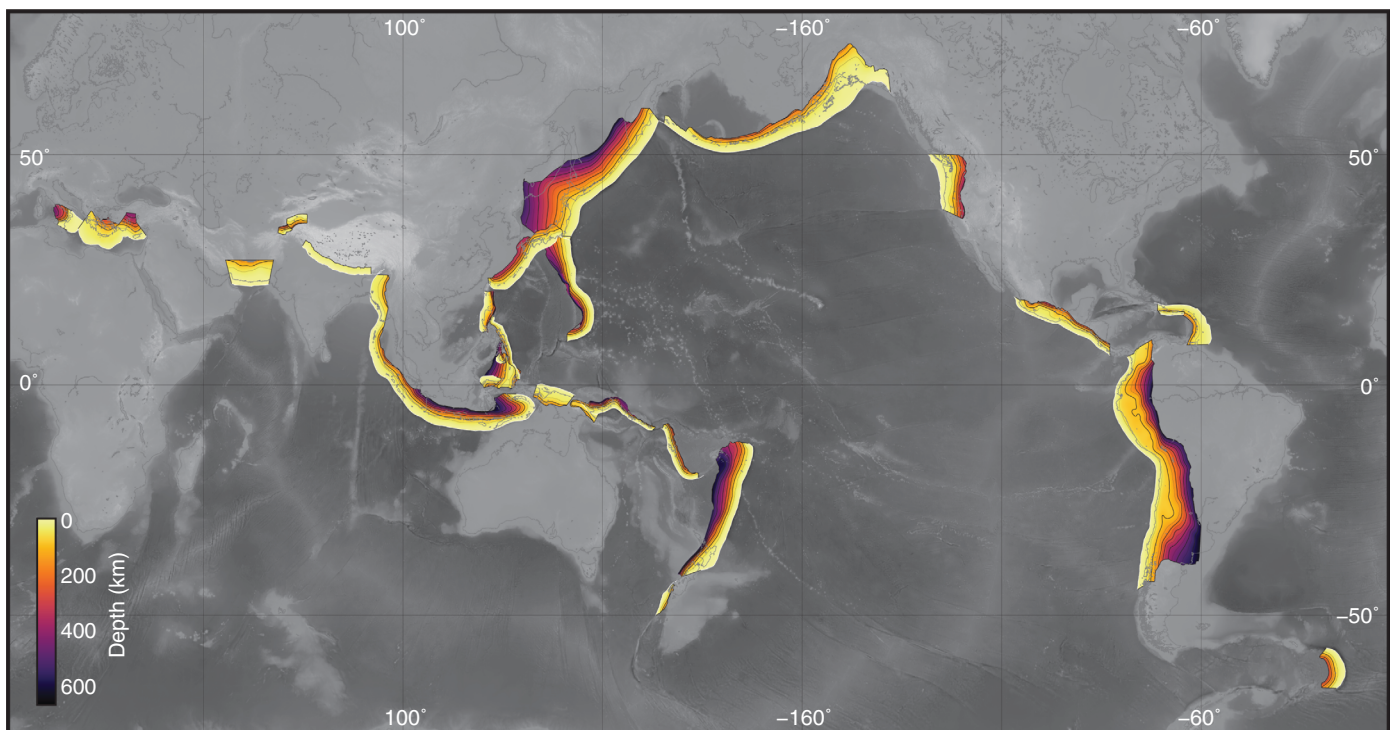
Fig. 1. The global distribution of models included in Slab2. Models are colored by depth. Slab2 and the code base used to create it are openly
available via ScienceBase (*31*).

an improved density of active-source seismic data interpretations, receiver functions, local and regional seismicity catalogs (from both regional networks and relocation studies), and seismic tomography models (*14*) (fig. S2). Our inclusion of these data facilitates a much more complete imaging of slab surfaces than is possible with global earthquake and moment tensor catalogs alone. Second, we have substantially modified our slab depth modeling approach. Rather than using the 2D-to-3D iterative process of Slab1.0, we moved straight to modeling the 3D structure of the slab. We first modeled the slab depth at individual nodes over a grid for each subduction zone (*15*) and then interpolated those depths onto a 3D grid, or surface.

Our approach (*15*) (figs. S3 to S5) allowed us to model all known active subduction zones globally, creating a complete collection of subduction zone geometry models, built routinely and via the same modeling approach (Table 1 and table S3). With Slab2, we improved how overturned slabs (e.g., the Izu-Bonin slab) are handled (fig. S6) by generating such slab surfaces in a rotated, strike-parallel reference frame. Slab2 includes models for sparsely seismic yet very hazardous subduction zones such as the Himalaya, Manila, Cotabato, Sulawesi, and northern New Guinea. Slab2 models the little-studied subduction zone geometries of the Halmahera and Papua New Guinea slabs, which both contain sections with inverted U-shaped structure (Figs. 2 and 3). In the Philippines region (Fig. 2), at least four subducting slabs overlap one another, creating a complex mosaic of subducting slabs in which the shallow Philippine Sea, Cotabato, and Sulawesi slabs are underlain by the large, buried, and inverted U-shaped Halmahera slab. Although the shape of the latter slab has been previously studied [e.g., (*16*) and (*17*)], the detailed interaction with surrounding slabs, including the smaller Cotabato and Sulawesi subduction zones, is poorly understood. Each of these slabs has hosted large to great-sized earthquakes in the past, and improved understanding of their geometries is critical for hazard analyses of the surrounding, highly populated region. Similarly, beneath Papua New Guinea, the Solomon Sea slab has subducted to the south and north (Fig. 3), as first observed by Ripper (*18*) and modeled more recently by Holm and Richards (*19*). Given the clear connectivity between the southern and northern limbs of this subduction zone seen in our geometry model and given that active subduction continues to the north, substantial tectonic complexity is implied in the transition region between the northward-subducting New Britain slab and the connected but buried arm of the slab farther west (Fig. 3).

Elsewhere, the addition of datasets improves how flat slab regions are defined. For example, in Central America, Slab1.0 failed to accurately capture the flat slab beneath southern Mexico because of its lack of seismicity. Additional receiver function datasets (*20*) allow Slab2 to better model the extent of the flat slab. This region is near the location of the destructive 19 September 2017 earthquake (*21*) (fig. S7). In South America,

tomographic datasets (*22*) enable Slab2 to resolve the full extent of the flat slab beneath southern Peru (fig. S8) and the aseismic deep slab beneath southern Chile. In the Cascadia region, tomographic data extend our image of the subducting slab well beyond the depth extent of previous models (e.g., Slab1.0) (*23*). In all cases, the expanded data coverage and improved methodology of Slab2 facilitate a more complete definition of subduction zone geometry than was possible with the model's predecessor, Slab1.0.

Inherent to our modeling approach is the creation of a filtered, slab-related catalog of seismicity for each subduction zone. Among other applications, these catalogs can be used to improve estimates of properties of the seismogenic zone (*1, 24*), commonly used for seismic hazard analyses. We define the shallow and deep limits of the seismogenic zone as the 10th and 90th percentiles of a double-normal distribution fit to the number of thrust earthquakes in our filtered dataset for each subduction zone. We modeled
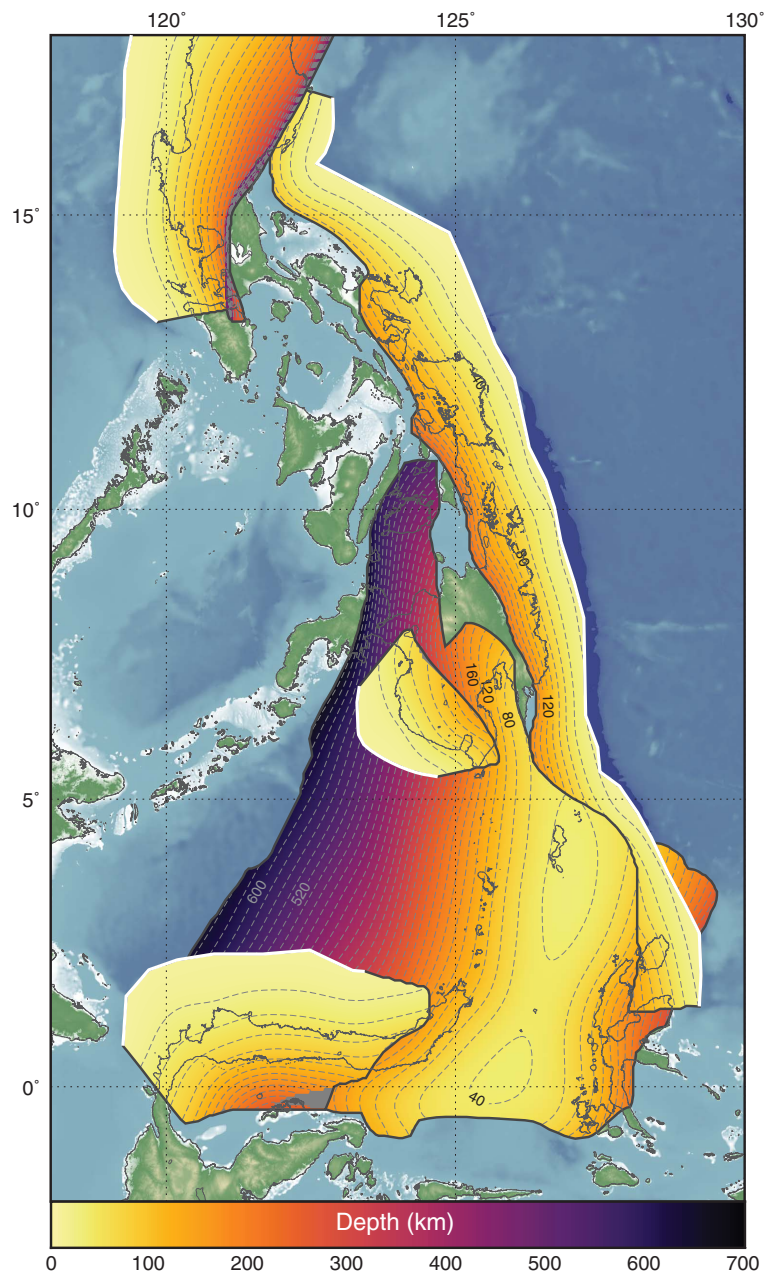
Fig. 2. Slab2 models in the region surrounding the Philippines. Here, five slabs (the Manila, Philippine, Cotabato, Halmahera, and Sulawesi slabs, from north to south) interact with one another. The Halmahera slab is an inverted U-shaped subduction zone that is completely buried and has no surface expression. It underlies the Cotabato and Sulawesi slabs on its western limb and abuts or underlies the Philippine slab on its eastern limb.

Fig. 3. Slab2 model for the Solomon Islands–Papua New Guinea slab.
(**A**) The inset shows global orientation. Here, the subducted slab transitions from east to west from a deep, northward-subducting body to one that has subducted in two directions, creating an inverted U-shaped structure beneath Papua New Guinea, as seen in the oblique view of the slab in (**B**), where the viewer is elevated above the surface of Earth to the northwest, looking southeast. In (**C**), we show a cross section of the slab along the vertical plane passing through the line A-A' in (A). Earthquakes (circles) and focal mechanisms are colored by their horizontal distance from the plane (red to the southeast, blue to the northwest of the plane; earthquakes within 50 km of the line are shown).

Table 1. A selection of subduction zone models included in Slab2. Shallow (Ss) and deep (Sd) seismogenic zone limits and corresponding seismogenic zone widths (Sw) are shown. See table S3 for expanded content. δ, Φ, and λ represent the average subduction zone interface dip, strike, and rake, respectively, in the seismogenic zone. n, number of contributing earthquakes; M_{max}, maximum historically observed magnitude for each subduction zone.

Model no.	Subduction zone arc	Slab2 code	n	Ss (km)	Sd (km)	Sw (km)	δ (°)	Φ (°)	λ (°)	M_{max}
1	Aleutians	alu	470	12	45	124	14	265	124	9.2
3	Central America	cam*	701	11	42	91	20	301	86	8.1
11	Izu-Bonin	izu	218	10	37	96	15	174	93	7.5
12	Kermadec	ker	707	10	47	129	16	205	93	8.3
13	Kuril	kur	1162	13	54	138	15	227	89	9.1
15	Manila	man	90	14	48	69	29	130	89	7.2
18	New Guinea	png	81	10	41	191	9	106	89	8.2
19	Philippines	phi	121	11	49	79	28	142	76	7.8
21	Ryukyu	ryu	265	13	47	111	17	236	99	8.8
22	South America	sam	1370	11	45	118	15	356	85	9.5
23	Scotia	sco	70	11	46	87	17	175	109	7.5
24	Solomon Islands	sol	627	12	53	77	33	287	96	8.0
26	Sumatra/Java	sum	706	11	53	153	17	281	100	9.1
27	Vanuatu	van	467	11	49	75	29	322	90	8.0

*In Slab1.0, this model was named "mex."
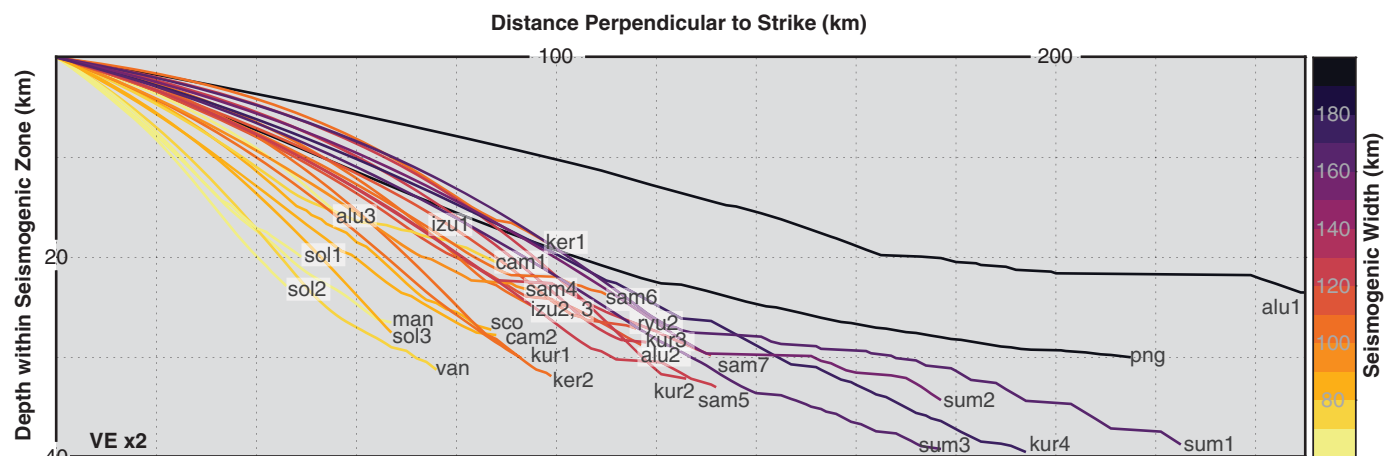
Fig. 4. The shape of the seismogenic zone in global subduction zones. Zones are color coded by the Sw. Line labels are defined in Table 1. The figure has a vertical exaggeration of 2 (VE ×2).

seismogenic zone width (Sw) (the length of the seismogenic zone along the dip direction of the slab) only for those slabs having more than 50 identifiable thrust-type earthquakes in our filtered datasets used to constrain slab depth (Fig. 4). The deep limit of the seismogenic zone varies globally from ~37 km (in the Izu-Bonin slab) to ~54 km (in the Kuril Islands). The seismogenic zone is nowhere more vertically broad than 40 km (measured as the difference between the shallow and deep limits). Sw varies from ~70 km (Manila slab) to ~195 km (Alaska slab), averaging about 115 km. Our estimates are ~30% larger than older estimates [e.g., (24)]. Instead of assuming a planar geometry with average subduction dip, we calculated Sw as the average down-dip distance between the upper and lower depth limits of the seismogenic zone along the nonplanar subduction zone interface. In Table 1, we also provide seismogenic zone properties for major sections within each subduction zone to better understand how seismogenesis varies along individual slabs.

When Sw is plotted versus the maximum observed earthquake magnitude in these subduction zones (fig. S9), we find a weak positive correlation. In general, the largest earthquakes occur where the seismogenic zone is wide (more than ~100 km), and narrow seismogenic zones (less than ~100 km) tend to host only small earthquakes, but notable outliers are evident. Southern Chile and Kamchatka have fairly narrow (Sw ~ 100 km) seismogenic zones yet have hosted magnitude 9+ earthquakes, whereas Java and New Guinea have extremely broad (Sw > 160 km) seismogenic zones but are not known to have hosted great-sized earthquakes. These results imply that Sw is not the singular controlling factor in potential earthquake size. Along-strike geometry (strike-parallel distance) is equally important (6), as are other seismogenic zone properties.

Slab2 is the most complete database of detailed 3D subduction zone geometries created. It includes slab geometries for the vast majority of global subduction zones and is built upon a robust and open-source methodology that should facilitate further improvements to our knowledge

of the fate of subducted slabs. There are a multitude of uses for these geometry models, from earthquake (4) and tsunami (10) hazard analyses to studies of mantle flow (11). As we have shown here, Sw does not play a dominant role in controlling the maximum size of earthquakes in subduction zones, but Slab2 provides the data necessary to better understand what geometrical control—if any—is exerted on earthquake rupture. To first order, Slab2 may also be used to identify whether segment boundaries in subduction zones are related to geometrical features on the megathrust (25), a key step in being able to assess the long-term persistence of subduction zone segmentation (26). These examples are just a few of many applications for Slab2, making it possible to better understand the interactions between slab bodies and a variety of other Earth processes.

REFERENCES AND NOTES

1. G. P. Hayes, D. J. Wald, R. L. Johnson, *J. Geophys. Res.* **117**, B01302 (2012).
2. G. P. Hayes *et al.*, *Nature* **512**, 295–298 (2014).
3. D. Melgar *et al.*, *Geophys. Res. Lett.* **43**, 961–966 (2016).
4. M. Pagani *et al.*, *Seismol. Res. Lett.* **85**, 692–702 (2014).
5. P. Audet, R. Bürgmann, *Nature* **510**, 389–392 (2014).
6. Q. Bletery *et al.*, *Science* **354**, 1027–1031 (2016).
7. T. H. Dixon *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 17039–17044 (2014).
8. C. M. Eakin, M. D. Long, *J. Geophys. Res.* **118**, 4794–4813 (2013).
9. G. M. Yogodzinski *et al.*, *J. Petrol.* **56**, 441–492 (2015).
10. E. Geist, P. Lynett, *Oceanography* **27**, 86–93 (2014).
11. M. Morishige, S. Honda, *Earth Planet. Sci. Lett.* **365**, 132–142 (2013).
12. O. Gudmundsson, M. Sambridge, *J. Geophys. Res.* **103** (B4), 7121–7136 (1998).
13. E. M. Syracuse, G. A. Abers, *Geochem. Geophys. Geosyst.* **7**, Q05017 (2006).
14. D. E. Portner, G. P. Hayes, *Geophys. J. Int.* **215**, 325–332 (2018).
15. Materials and methods are available as supplementary materials.
16. R. Hall, *Tectonophysics* **144**, 337–352 (1987).
17. J. Wu, J. Suppe, R. Lu, R. Kanda, *J. Geophys. Res.* **121**, 4670–4741 (2016).
18. I. D. Ripper, *Tectonophysics* **87**, 355–369 (1982).
19. R. J. Holm, W. Richards, *Aust. J. Earth Sci.* **60**, 605–619 (2013).
20. X. Pérez-Campos *et al.*, *Geophys. Res. Lett.* **35**, L18303 (2008).
21. D. Melgar *et al.*, *Geophys. Res. Lett.* **45**, 2633–2641 (2018).
22. A. Scire, G. Zandt, S. Beck, M. Long, L. Wagner, *Geosphere* **13**, 665–680 (2017).
23. P. A. McCrory, J. L. Blair, F. Waldhauser, D. H. Oppenheimer, *J. Geophys. Res.* **117**, B09306 (2012).
24. J. F. Pacheco, L. R. Sykes, C. H. Scholz, *J. Geophys. Res.* **98** (B8), 14133–14159 (1993).
25. G. P. Hayes, S. M. Plescia, G. Moore, "Inferring rupture characteristics using new databases for 3D slab geometry and earthquake rupture models," paper presented at the 2017 Fall Meeting of the American Geophysical Union, New Orleans, LA, 11 to 15 December 2017 (T23F-0663).
26. A. J. Meltzner *et al.*, *J. Geophys. Res.* **117**, B04405 (2012).
27. U.S. Geological Survey, Search Earthquake Catalog; http://earthquake.usgs.gov/earthquakes/search/.
28. E. R. Engdahl, R. D. Van Der Hilst, R. P. Buland, *Bull. Seismol. Soc. Am.* **88**, 722–743 (1998).
29. G. Ekström, M. Nettles, A. M. Dziewoński, *Phys. Earth Planet. Inter.* **200–201**, 1–9 (2012).
30. P. Wessel, W. H. F. Smith, *EOS Trans. Am. Geophys. Union* **72**, 441–446 (1991).
31. G. P. Hayes, The Slab2 online repository, ScienceBase (2018); https://doi.org/10.5066/F7PV6JNV.

SUPERCONDUCTIVITY

Rapid change of superconductivity and electron-phonon coupling through critical doping in Bi-2212

Y. He[1,2]*, M. Hashimoto[3]*, D. Song[4]†, S.-D. Chen[1,2], J. He[1,2], I. M. Vishik[1]‡, B. Moritz[1,2], D.-H. Lee[5], N. Nagaosa[6], J. Zaanen[7], T. P. Devereaux[1,2], Y. Yoshida[4], H. Eisaki[4], D. H. Lu[3], Z.-X. Shen[1,2]§

Electron-boson coupling plays a key role in superconductivity for many systems. However, in copper-based high–critical temperature (T_c) superconductors, its relation to superconductivity remains controversial despite strong spectroscopic fingerprints. In this study, we used angle-resolved photoemission spectroscopy to find a pronounced correlation between the superconducting gap and the bosonic coupling strength near the Brillouin zone boundary in $Bi_2Sr_2CaCu_2O_{8+\delta}$. The bosonic coupling strength rapidly increases from the overdoped Fermi liquid regime to the optimally doped strange metal, concomitant with the quadrupled superconducting gap and the doubled gap-to-T_c ratio across the pseudogap boundary. This synchronized lattice and electronic response suggests that the effects of electronic interaction and the electron-phonon coupling (EPC) reinforce each other in a positive-feedback loop upon entering the strange-metal regime, which in turn drives a stronger superconductivity.

The phase diagram of cuprate high-temperature superconductors hosts a number of complex orders, types of fluctuations, and interactions (1–4). In the non–Fermi liquid strange-metal regime, a hierarchy of microscopic interactions are intimately at play but not fully understood (1, 2, 4). Although the experimental evidence for d-wave superconductivity (5–7) naturally points to an electron-electron interaction–based pairing mechanism (8–12), the omnipresent charge order (3) points to the role of electron-phonon coupling (EPC), especially in the context of enhanced EPC by electronic correlation (13, 14) and multichannel boosted superconductivity (15–17). Although there have been reports of EPC imprinting on the electronic structure of many cuprate superconductors (18–21), little evidence directly correlates EPC with the intertwined orders in the phase diagram (1, 2). By focusing on the overdoped side in $Bi_2Sr_2CaCu_2O_{8+\delta}$ (Bi-2212), we used angle-resolved photoemission spectroscopy (ARPES) to uncover a set of pronounced effects rapidly crossing over from the overdoped Fermi liquid regime to the optimally doped strange

metal, closely associated with the putative pseudogap quantum critical point.

To understand the doping dependence of the superconducting character, we first investigated the energy gap. Figure 1A depicts the recently proposed phase diagram, with an emphasis on the pseudogap phase boundary (red boxed area) (22, 23). In this doping range, the system is marked by a strange-metal normal state characterized by spectral incoherence and linear resistivity at high temperatures (2), followed by intertwined states with pseudogap and superconductivity at low temperatures (23). Figure 1B plots the Fermi-Dirac function divided energy distribution curves (FD-EDCs) at the antinodal Fermi momentum k_F from hole doping (per Cu atom) $p = 0.14$ [underdoped critical temperature (T_c) = 92 K, UD92] to $p = 0.24$ (overdoped T_c = 47 K, OD47), for both low temperatures (blue; $T \ll T_c$) and high temperatures [red; $T \sim T^*$ (where T^* is the pseudogap temperature) for pseudogap regime, $T > T_c$ for nonpseudogap regime, generally referred to as $T \gg T_c$ hereafter] (24)]. The spectra in the deeply overdoped region are characterized by prominent Bogoliubov quasiparticle peaks at low temperature ($T \ll T_c$) and pronounced coherent peaks in the normal state ($T \gg T_c$). The extracted antinodal gap size Δ_{AN} at $T \ll T_c$ is plotted as a function of the superconducting transition temperature in Fig. 1C, revealing a fourfold change in contrast to the only twofold change in T_c across the doping range of interest. The gap-to-T_c ratio reaches the weak coupling d-wave Bardeen-Cooper-Schrieffer (BCS) value in the doping range where $T_c \lesssim 65$K [lower dashed line at $2\Delta/k_B T_c = 4.3$ (where k_B is the Boltzmann constant)], and the scattering dynamics can be described by Fermi liquid theory (fig. S2). Figure 1D shows the doping dependence of the

gap-to-T_c ratio $2\Delta/k_B T_c$, which rapidly increases from ~4.3 at $p \sim 0.22$, a doping where $T_c \sim T^*$ (22, 23), to ~10 at $p \sim 0.19$, a doping where a pseudogap critical point has been suggested (2, 25). In this rapid process of the superconducting character change, T_c rises from 65 to 95 K despite the increasing presence of the competing pseudogap (23). Throughout this doping range, the low-temperature spectral gap takes the d-wave form, which means that the antinodal gap size Δ_{AN} embodies the entire d component (also known as the nodal gap Δ_N) in this discussion (fig. S3) (22, 24). This doping range is to be distinguished from the underdoped region, where the energy gap deviates from the d-wave form (4, 22). Indeed, Δ_{AN} and Δ_N show consistent behavior (Fig. 1, C and D).

With the rapid doping evolution of the superconducting character, the EPC around the Brillouin zone (BZ) boundary also undergoes a substantial change. Figure 2A shows the FD-EDCs from (π, 0), where the peak directly reflects the antinodal band bottom (high temperatures, red), and the spectral minima in the "peak-dip-hump" structure are minimally disturbed by the bonding band (low temperatures, blue) (24). Such spectral dips are commonly interpreted as the fingerprint of electron-boson coupling and become prominent with strong EPC (16, 18, 24, 26). This is inferred from the variation with doping of the difference between low- and high-temperature spectra (highlighted in gray); the spectral weight of this difference increases with growing EPC (Fig. 2B and fig. S4). Note that the effect of the weakly temperature dependent bilayer splitting is minimized in this analysis (24). As shown in Fig. 2D and fig. S6B, the dip position tracks the rapid evolution of Δ_{AN} at $T \ll T_c$ by an energy offset of ~37 meV, which is nearly independent of doping, agrees well with the energy of the B_{1g} oxygen bond-buckling phonon mode (18, 24), and is qualitatively consistent with previous scanning tunneling microscope studies (20). Such a mode was also revealed in recent high-resolution photoemission studies on similarly overdoped systems near the antinode (27). Surprisingly, the spectral weight associated with the dip, which reflects the EPC strength (24), abruptly grows between $p \sim 0.22$ and 0.19 (Fig. 2E). This finding, along with the similar behavior in doping-dependent charge-transport properties (28), cannot be explained by a simple chemical potential shift, which has yet to cross the band bottom at (π, 0) (Fig. 2A). For comparison, the predepletion of the density of states (DOS) from the high-temperature normal state to T_c (ratio between the red and blue arrows in Fig. 2C; also see fig. S6C) is plotted in Fig. 2F. Compared with the conventionally used T^* or Δ_{AN}, this quantity better disentangles the pseudogap contribution from the bulk superconductivity, reflecting the direct DOS depletion effects dominated by the pseudogap. Clearly, all three quantities—superconductivity (Fig. 2D), pseudogap (Fig. 2F), and EPC strength (Fig. 2E)—show concomitant onset and rapid change in the interval $p = 0.19$ to 0.22 in a highly correlated fashion.

[1]Geballe Laboratory for Advanced Materials, Departments of Physics and Applied Physics, Stanford University, Stanford, CA 94305, USA. [2]SIMES, SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA. [3]Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA. [4]National Institute of Advanced Industrial Science and Technology, Tsukuba 305-8568, Japan. [5]Department of Physics, University of California, Berkeley, CA 94720, USA. [6]Quantum-Phase Electronics Center, Department of Applied Physics, University of Tokyo, Tokyo 113-8656, Japan. [7]Instituut-Lorentz for Theoretical Physics, Leiden University, Leiden, Netherlands.
*These authors contributed equally to this work. †Present address: Department of Physics and Astronomy, Seoul National University, Seoul 08826, Republic of Korea. ‡Present address: Department of Physics, University of California, Davis, CA 95616, USA.
§Corresponding author. Email: zxshen@stanford.edu

By showing plots of integrated spectra near the antinode, Figure 2C highlights the marked contrast between the two distinct regimes: the complex strange-metal regime near optimal doping (C1; strong pairing, strong EPC, and pseudogap) and the deeply overdoped region, in which the system behavior appears to submit to the classic BCS rules (C2; weaker pairing, weaker EPC, and no pseudogap). The contrast in the spectra is most pronounced in the nor-mal state where no quasiparticles exist in the strange-metal state, but Fermi liquid quasi-particles are apparent in the deeply overdoped regime. The main discovery is that the strange-metal phase involves not only strong correlations and incipient intertwining electronic orders but also lattice degree of freedom (in particular, the B_{1g} phonon for this system). The peak-dip-hump structure is the low-temperature remnant of such intertwinement (after superconductivity has already set in) and is largely smeared out at high temperatures by the spectral incoherence characteristic of the strange-metal state.

On the basis of the doping-dependent inves-tigation (Fig. 2, D to F), we suggest that the in-creasing electronic correlation, marked by the entrance into the strange-metal regime and the emergence of pseudogap, may have triggered this sudden EPC enhancement and the com-plex feedback interaction that follows. It is

Fig. 1. Superconductivity rapidly deviates from the weakly coupled d-wave BCS limit with underdoping. (**A**) Schematic phase diagram for Bi-2212 [reproduced from (22)], with an emphasis on the doping range of interest (boxed in red). Blue squares represent the antinodal gap-opening temperature T^*; red circles denote the superconducting transition temperature T_c (fig. S1). PG, pseudogap; SC, superconductivity; SM, strange metal; FL, Fermi liquid. (**B**) FD-EDCs taken at the antinodal k_F for a series of doping. Red curves refer to $T \sim T^*$ for $p < 0.22$ and $T > T_c$ for $p > 0.22$, and blue curves indicate $T \ll T_c$ for all dopings (24). Black circles denote the superconducting quasiparticle peak position. The inset schematically illustrates the momenta of the listed spectra. arb. u., arbitrary units; E, energy; E_F, Fermi energy. (**C**) Antinodal gap Δ_{AN} (solid circles) and nodal gap Δ_N (open circles) (22) plotted against the sample T_c. The two dashed lines mark the gap-to-T_c ratios of 4.3 (the weakly coupled d-wave BCS value) and 10. (**D**) The gap-to-T_c ratio shows a rapid deviation from 4.3 (horizontal black dashed line) at $p < 0.22$. In (A), (C), and (D), error bars are propagated from T_c uncertainties (see fig. S1).
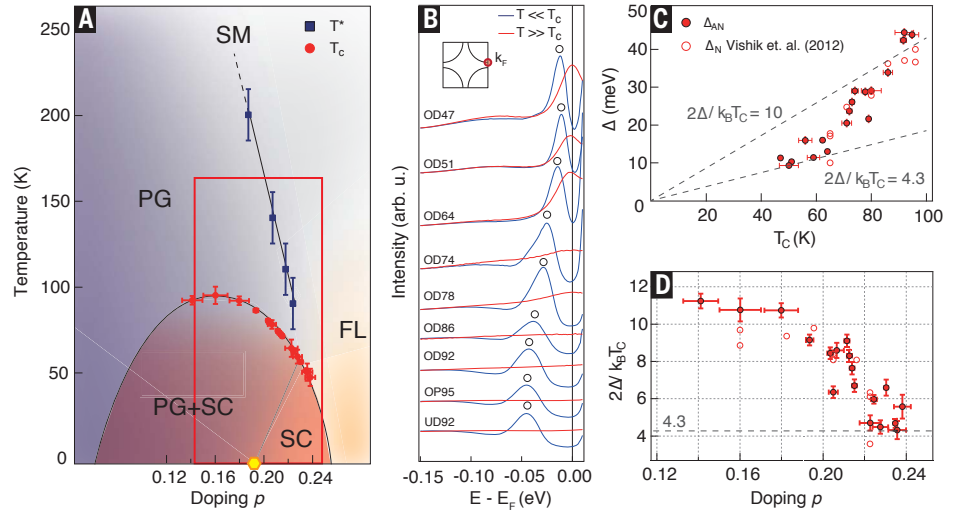
Fig. 2. Abrupt growth of the EPC strength with underdoping. (**A**) The $(\pi, 0)$ FD-EDCs taken at low temperature (blue) and high temperature (red, $T \sim T^*$ for pseudogap regime, $T > T_c$ for sufficiently overdoped regime) from heavy overdoping (OD47) to slight underdoping (UD92). (**B**) Spectral intensity difference between the two temperatures for each doping. Black circles denote the dip maxima. (**C**) Integrated FD-EDC over the antinodal momenta at low temperatures ($T \ll T_c$, blue), just above T_c ($T \sim T_c^+$, green), and high temperatures ($T \gg T_c$, red) for the $p = 0.16$ sample (OP95, C1) and the $p = 0.23$ sample (OD51, C2). The momentum integration range is noted by the red bar in the inset schematic BZ. The spectra are normalized to the total spectral weight between binding energies of 200 and 300 meV. (**D**) Doping-dependent dip position and coevolving antinodal gap Δ_{AN}. (**E**) Doping-dependent spectral weight (SW) of the dip [gray area in (B)] relative to the total spectral weight of the normal state. (**F**) Doping-dependent antinodal spectral weight depletion at T_c^+ (red arrow in C1) relative to the full depletion at $T \ll T_c$ (blue arrow in C1), derived from the intensity of the integrated FD-EDCs within ±2 meV of E_F. In (D) to (F), error bars are propagated from T_c uncertainties (see fig. S1).
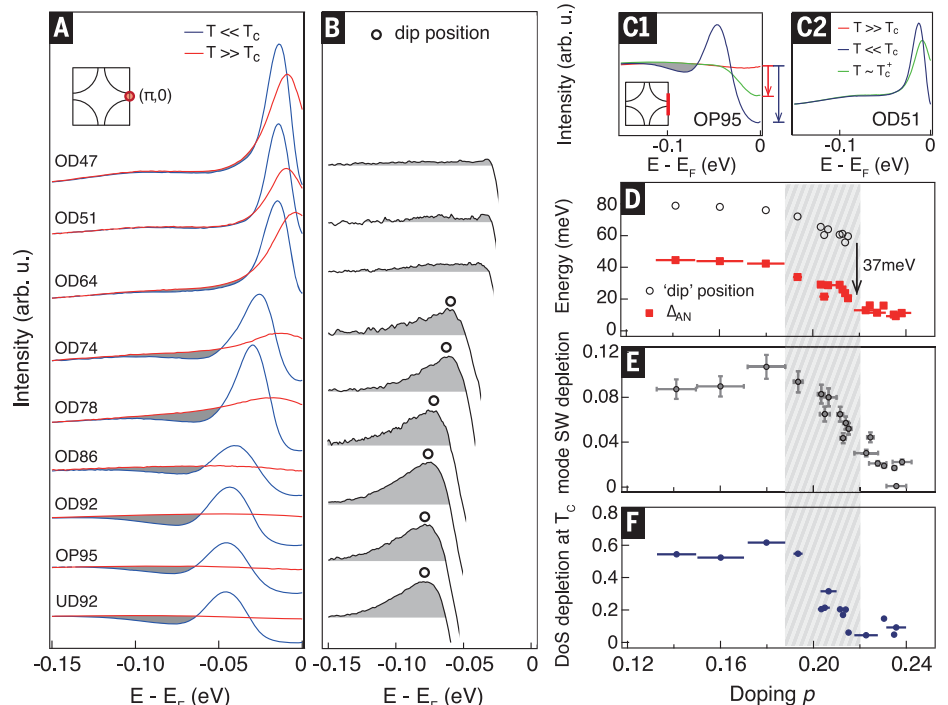
Fig. 3. Intertwined growth of super-conductivity and EPC tuned by the hole concentration. The red line is an illustration of the T_c in Bi-2212 (T_c^{max} = 95 K). The blue shaded region and line represent the single-layer Bi-2201 system, in which the coupling to the B_{1g} mode is weak and T_c^{max} is only 38 K. The yellow ball represents the optimally doped trilayer Bi-2223, in which T_c^{max} is 108 K. The top-right inset shows the intertwined relation between the pseudogap and the EPC under strong electronic correlation. The Madelung potential and the lattice stacking along the c axis are schematically depicted for the single-layer, bilayer, and trilayer systems (left insets). The dark gray blocks represent the CuO_2^{n-} plane, and the light gray blocks represent the charge reservoir layers (Ca^{2+}, SrO, BiO^+). The orange circles mark the CuO_2^{n-} planes that experience to the first order a nonzero out-of-plane electric field.
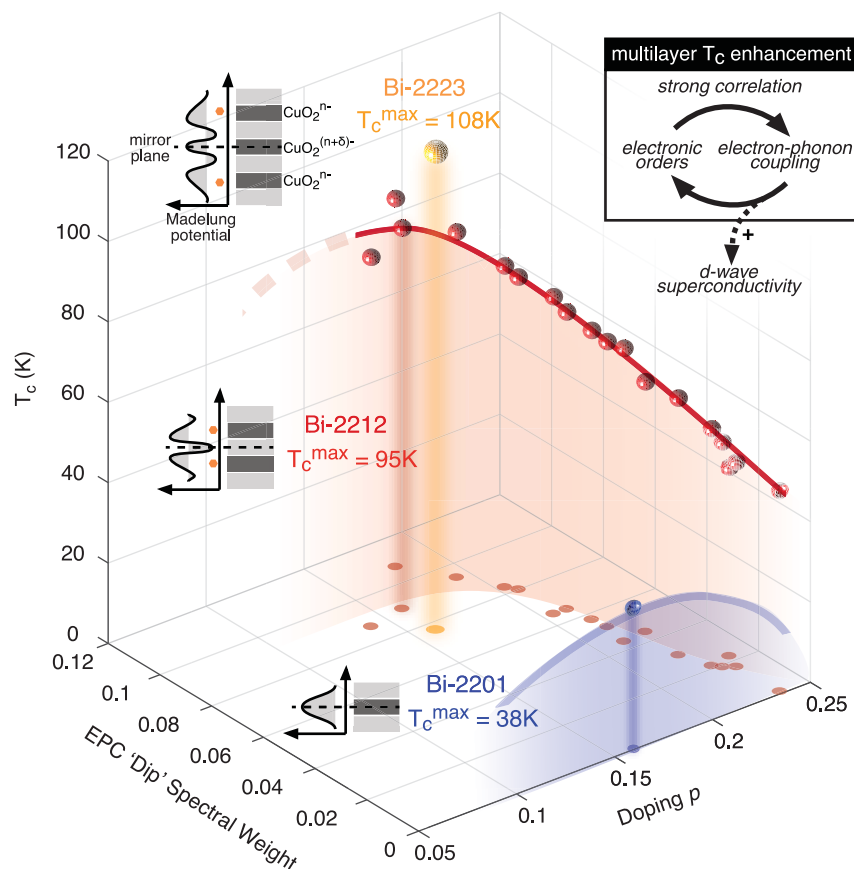
likely that between p = 0.22 and 0.19 the charge carriers slow down, the charge screening becomes inoperative, and the strange metal takes over (also see fig. S8). With further underdoping and stronger pairing, EPC and the pseudogap become closely tied phenomena (Fig. 3, top right inset). Although the electronically driven charge density wave (CDW) does not result in substantial static lattice deformation, the "dynamical" effects of EPC—giant phonon anomalies (*29*), the large sensitivity of the charge order and superfluid density to isotope substitution (*19*), as well as the strengthening of the superconductivity caused by phonon pumping (*30*)—are indeed abundant in the underdoped region. Our present results indicate that the non–Fermi liquid nature of the strange metal appears to have a drastic impact on the fundamental physics of the EPC, which may well be at the origin of these highly anomalous EPC effects on the intertwined orders at low temperatures (*31*). Consequently, EPC and strange-metal physics play together to enhance the system's tendency to order, inducing instability that leads to CDW or more complex electronic textures (*3*). Further underdoping leads to extreme EPC amplified by correlation effects, where the charge carriers eventually develop polaronic signatures (*13*). In particular, the B_{1g} phonon is suggested to accommodate the d-wave form factor of both the CDW (particle-hole channel) and the superconductivity (particle-particle channel) in the

cuprate systems, providing pathways for the lattice to interact with the electronic orders (*32*).

Indeed, the pseudogap temperature T^* near optimal doping receives a notable increase from single-layer to bilayer Bi-based cuprate systems (fig. S8C). Only in the multilayer case does the oxygen B_{1g} mode substantially couple to electrons: As the CuO_2^{n-} planes do not lie on a crystal mirror plane, a coupling can only arise from first-order c axis atomic displacements (Fig. 3, left insets). Such layer dependence also holds when the trilayer system is taken into account (*33*). Therefore with underdoping, the rapid formation of the pseudogap and enhanced EPC could be perceived as a result of a positive feedback under increasing electronic correlation, which efficiently reinforces the system's already strong tendency to order.

More intriguingly, superconductivity may benefit from this feedback loop. Figure 3 compiles the superconducting T_c of the single-layer (Bi-2201, blue), bilayer (Bi-2212, red), and trilayer (Bi-2223, yellow) systems as a function of doping and antinodal EPC dip strength. The maximum T_c sees a considerable increase from single-layer to multilayer systems. Although the growing electronic correlation strength from the overdoped regime toward the optimal doping may have strengthened the pseudogap (*1–4*), it also joins efforts with the enhanced EPC to achieve the higher T_c in multilayer cuprate superconductors than in their single-layer counterparts

(*33*). As a corroborative example, a recent photoemission study on a monolayer FeSe film system demonstrated the experimental viability to have a secondary phononic channel as the superconductivity enhancer (*16*), boosting T_c by as much as 50% in comparison with the uncoupled case. The oxygen bond-buckling B_{1g} phonon couples strongly in the multilayer cuprates, owing to the steep Madelung potential going through the CuO_2^{n-} planes that do not constitute the crystal mirror plane (Fig. 3, left insets, orange circles). This phonon mode is also known to favor scattering with small momentum transfer, owing to its preferential coupling to electrons in the antinodal region, and is inherently poised for d-wave superconductivity enhancement (*15*). This complex involvement of EPC in the pairing process may also be the underlying reason for the dynamical and diverse manifestation of isotope effects (*19, 20*). Therefore, it may be insufficient to regard Cooper pairing as driven by pure electron-electron interactions in the optimal and underdoped regions.

On the other hand, the deeply overdoped regime emerges as a weakly coupled, simpler superconductor with a respectably high T_c in excess of 50 K, which offers a cleaner platform to investigate the pairing mechanism. Meanwhile, with enhanced correlation in the strange-metal regime, EPC and the intertwined electronic orders form a positive-feedback loop (Fig. 3, right inset, solid arrows) that ultimately provides an

additional pathway for T_c enhancement by EPC in the multilayer systems (dashed arrow). From a broader perspective, cuprate superconductors, given this interpretation of their rich phase diagram, may be generalized as a new model system for testing EPC-enhanced superconductivity through multiple underlying channels. Systematic investigation of the B_{1g} phonon self energy around the critical doping via various scattering techniques may provide valuable insights into such scenarios.

REFERENCES AND NOTES

1. E. Fradkin, S. A. Kivelson, J. M. Tranquada, *Rev. Mod. Phys.* **87**, 457–482 (2015).
2. B. Keimer, S. A. Kivelson, M. R. Norman, S. Uchida, J. Zaanen, *Nature* **518**, 179–186 (2015).
3. R. Comin, A. Damascelli, *Annu. Rev. Condens. Matter Phys.* **7**, 369–405 (2016).
4. M. Hashimoto, I. M. Vishik, R. H. He, T. P. Devereaux, Z.-X. Shen, *Nat. Phys.* **10**, 483–495 (2014).
5. W. N. Hardy, D. A. Bonn, D. C. Morgan, R. Liang, K. Zhang, *Phys. Rev. Lett.* **70**, 3999–4002 (1993).
6. Z. Shen et al., *Phys. Rev. Lett.* **70**, 1553–1556 (1993).
7. D. A. Wollman, D. J. Van Harlingen, W. C. Lee, D. M. Ginsberg, A. J. Leggett, *Phys. Rev. Lett.* **71**, 2134–2137 (1993).
8. D. J. Scalapino, E. Loh Jr., J. E. Hirsch, *Phys. Rev. B* **34**, 8190–8192 (1986).
9. G. Baskaran, Z. Zou, P. W. Anderson, *Solid State Commun.* **63**, 973–976 (1987).
10. G. Kotliar, J. Liu, *Phys. Rev. B* **38**, 5142–5145 (1988).
11. P. A. Lee, N. Nagaosa, X.-G. Wen, *Rev. Mod. Phys.* **78**, 17–85 (2006).
12. P. Monthoux, D. Pines, G. G. Lonzarich, *Nature* **450**, 1177–1183 (2007).
13. A. S. Mishchenko, N. Nagaosa, *Phys. Rev. Lett.* **93**, 036402 (2004).
14. S. Gerber et al., *Science* **357**, 71–75 (2017).
15. S. Johnston et al., *Phys. Rev. B* **82**, 064513 (2010).
16. J. J. Lee et al., *Nature* **515**, 245–248 (2014).
17. Z.-X. Li, F. Wang, H. Yao, D.-H. Lee, *Sci. Bull.* **61**, 925–930 (2016).
18. T. Cuk et al., *Phys. Stat. Sol. B* **242**, 11–29 (2005).
19. J. L. Tallon, R. S. Islam, J. Storey, G. V. M. Williams, J. R. Cooper, *Phys. Rev. Lett.* **94**, 237002 (2005).
20. J. Lee et al., *Nature* **442**, 546–550 (2006).
21. H. Anzai et al., *Sci. Rep.* **7**, 4830 (2017).
22. I. M. Vishik et al., *Proc. Natl. Acad. Sci. U.S.A.* **109**, 18332–18337 (2012).
23. M. Hashimoto et al., *Nat. Mater.* **14**, 37–42 (2015).
24. Materials and methods are available as supplementary materials.
25. S. Sachdev, *Phys. Stat. Sol. B* **247**, 537–543 (2010).
26. D. J. Scalapino, J. R. Schrieffer, J. W. Wilkins, *Phys. Rev.* **148**, 263–279 (1966).
27. J. He, in *Angle-Resolved Photoemission Spectroscopy on High-Temperature Superconductors: Studies of Bi-2212 and Single-Layer FeSe Film Grown on SrTiO₃ Substrate* (Springer, 2016), pp. 71–79.
28. Y. Ando, Y. Kurita, S. Komiya, S. Ono, K. Segawa, *Phys. Rev. Lett.* **92**, 197001 (2004).
29. M. Le Tacon et al., *Nat. Phys.* **10**, 52–58 (2014).
30. R. Mankowsky et al., *Nature* **516**, 71–73 (2014).
31. J. Zaanen, Y. Liu, Y.-W. Sun, K. Schalm, in *Holographic Duality in Condensed Matter Physics* (Cambridge Univ. Press, 2015), chap. 2.
32. C. Honerkamp, H. C. Fu, D.-H. Lee, *Phys. Rev. B* **75**, 014503 (2007).
33. T. Sato et al., *Phys. Rev. Lett.* **89**, 067005 (2002).

SOLID-STATE PHYSICS

Quantum oscillations of electrical resistivity in an insulator

Z. Xiang[1], Y. Kasahara[2], T. Asaba[1], B. Lawson[1,3], C. Tinsman[1], Lu Chen[1], K. Sugimoto[4], S. Kawaguchi[4], Y. Sato[2], G. Li[1], S. Yao[5], Y. L. Chen[6], F. Iga[7], John Singleton[8], Y. Matsuda[2]*, Lu Li[1]*

In metals, orbital motions of conduction electrons on the Fermi surface are quantized in magnetic fields, which is manifested by quantum oscillations in electrical resistivity. This Landau quantization is generally absent in insulators. Here, we report a notable exception in an insulator—ytterbium dodecaboride (YbB$_{12}$). The resistivity of YbB$_{12}$, which is of a much larger magnitude than the resistivity in metals, exhibits distinct quantum oscillations. These unconventional oscillations arise from the insulating bulk, even though the temperature dependence of the oscillation amplitude follows the conventional Fermi liquid theory of metals with a large effective mass. Quantum oscillations in the magnetic torque are also observed, albeit with a lighter effective mass.

I n Kondo insulators, the hybridization between itinerant and localized electrons opens an insulating gap, and consequently, their resistivity diverges at low temperature (*1*, *2*). Recently, there has been a lively debate about the nature of the ground state of Kondo insulator samarium hexaboride (SmB$_6$) in intense magnetic fields. Although mixed-valence SmB$_6$ is a good insulator (its resistance increases by five orders of magnitude when cooled down to 300 mK from room temperature), Landau-level (LL) quantization still occurs, and clear quantum oscillations are observed in its magnetization [the de Haas–van Alphen (dHvA) effect] (*3*, *4*). The origin and interpretation of the dHvA oscillations in SmB$_6$ have been highly controversial, owing to a number of peculiar features. First, the oscillations are observed only in magnetization and not in electrical resistivity [the Shubnikov–de Haas (SdH) effect]. Second, unlike the heavy carriers revealed by the thermoelectric studies (*5*), oscillations appear to arise from quasiparticles with a very light effective mass ($m \ll m_e$, where m_e is the free electron mass) (*3*). Third, in floating zone–grown SmB$_6$ samples, the dHvA signal exhibits a striking deviation from the standard Lifshitz-Kosevich (LK) formula in Fermi-liquid theory (*4*). These observations point to either a topologically protected surface state (*3*) or the presence of an unconventional Fermi surface in an insulator (*4*). A number of

intriguing physical origins have been proposed, including exciton-based magnetic breakdown (*6*, *7*), Majorana-type charge-neutral Fermi surfaces (*8*), a failed superconducting ground state (*9*), and spinon Fermi surfaces (*10*, *11*). A key to solving the most fundamental problem, the existence of "a Fermi surface in an insulator," lies in clarifying whether quantum oscillations—in particular, in charge transport—are observable in another insulating system.

Here, we present quantum oscillation studies of ytterbium dodecaboride (YbB$_{12}$), another cubic-structured rare-earth intermetallic compound. YbB$_{12}$ has long been known as a mixed-valence Kondo insulator (*12*–*14*). It behaves as a monovalent metal with localized magnetic moments at room temperature, whereas a nonmagnetic insulating ground state develops at low temperatures. The opening of a narrow energy gap of 10 to 15 meV at the Fermi level has been confirmed with many experiments (*15*–*17*) and is attributed to the hybridization between the itinerant $5d$ and the localized $4f$ electrons. The mean valence of Yb ions in YbB$_{12}$ is +2.9 (*18*), close to $4f^{13}$ (+3) configuration. Therefore, the f electrons are mostly localized, and the crystalline electric field (CEF) ground state is well defined. By contrast, because the mean valence of Sm ions in SmB$_6$ is +2.6 (*19*), f electrons are more itinerant, and the CEF scheme is not well defined. Therefore, the electronic structure of YbB$_{12}$ is much simpler than that of SmB$_6$. Furthermore, early calculations in YbB$_{12}$ predicted the existence of topological surface states owing to mirror-symmetry protection (*20*).

YbB$_{12}$ single crystals are grown in a floating-zone furnace [(*21*), section 1]. Using the experimental setup shown in Fig. 1A, magnetic torque and magnetoresistance (MR) are measured up to 45 T simultaneously [(*21*), section 2]. The temperature dependence of the resistivity (Fig. 1B) confirms an increase of five orders of magnitude from room temperature to 50 mK. The resistivity has a weak temperature dependence below 2.2 K, resem-

bling the resistive "plateau" well known in SmB$_6$ at temperature (T) < 3.5 K (*22*). This "plateau" is an indication of the existence of extended in-gap states. Fitting with the thermal activation model of resistivity, $\rho(T) = \rho_0 \exp(\Delta/2k_BT)$, where k_B is the Boltzmann constant, reveals a two-gap feature with the gap width 12.5 meV (20 K < T < 40 K) and 4.7 meV (6 K < T < 12.5 K), respectively (Fig. 1B, inset), which is consistent with previous transport results (*23*). Upon applying the magnetic field, the negative slope of the $\rho(T)$ curve is preserved up to 45 T, with no hints of metallic behavior (Fig. 1C), indicating that the ground state is still gapped (fig. S5).

The field dependence of the magnetic torque in the insulating state of YbB$_{12}$ is shown in Fig. 1D. We observed a step increase at 20 T followed by a decrease at 28 T. These features are weak metamagnetic transitions and/or crossovers that could potentially be related to the predicted field-induced staggered magnetism in Kondo insulators (*24*). Above ~37.5 T, the dHvA oscillations are clearly resolved (Fig. 1D). The dHvA oscillations appear well below the insulator-metal (I-M) transition field in our YbB$_{12}$ samples, which was determined by means of pulsed field studies to be 45.3 to 47.0 T [(*21*), sections 4 and 5]. Fast Fourier transform (FFT) on the $\phi = 11.3°$ torque curve gives a dHvA frequency of $F = 720$ T (Fig. 1D, inset).

In Fig. 2A, the MR data at $\phi = 27.7°$ is plotted between 11.5 and 45 T. Given the zero-field resistivity $\rho(0) = 4.67$ ohm · cm of this sample at 350 mK, a significant negative MR {[$\rho(H) – \rho(0)$]/$\rho(0)$} of –95.9% is achieved at 45 T (a detailed angular dependence of MR is shown in fig. S3). The negative MR is a hallmark of Kondo insulators that results from field suppression of the hybridization gap (*25*–*28*). The negative MR in YbB$_{12}$ is much larger than that in SmB$_6$ (*29*, *30*). This is probably a consequence of the larger effective Landé g-factor in YbB$_{12}$, which increases the influence of the magnetic field (*17*, *28*). As the magnetic field (H) increases, MR displays wiggle-like features at around 16 and 28 T. These wiggles do not arise from quantum oscillations because their positions are obviously temperature dependent (fig. S7); instead, they are likely to be field-induced transitions or crossovers. The feature at 28 T is probably linked to the kink feature in the magnetic torque [(*21*), section 6].

The most striking result in Fig. 2A is the oscillations that appear in the MR under strong magnetic fields. From 40.8 up to 45 T, two valleys and one peak in total can be clearly observed (Fig. 2A, top inset). The FFT on the MR oscillations in this field regime yields a clear frequency peak at $F = 913$ T (Fig. 2A, bottom inset). With the magnetic field direction close to the crystal axes, up to four oscillation periods can be seen (Fig. 2B). The overall SdH patterns are almost identical for $\phi = 11.3°$ and $\phi = 78.4°$, whereas there is a small valley position shift between $\phi = 18.3°$ and $\phi = 70.5°$. This suggests an axis of symmetry along [101] direction for the SdH oscillations, which is consistent with the cubic structure of YbB$_{12}$ crystal (fig. S10). The valleys in $d\rho/dH$ in Fig. 2B being approximately uniformly

spaced as a function of $1/H$ provides strong evidence that the SdH oscillations have a single dominating frequency. These overall patterns of the SdH oscillations, as well as those of the dHvA oscillations, are well reproduced between different YbB$_{12}$ samples [(21), section 3].

The observation of the SdH oscillations is reinforced by the temperature dependence of oscillation amplitudes. The evolution of the high-field wiggle features in both MR and torque data shows the typical behavior of quantum oscillations, with temperature-independent positions of the dominant peaks and valleys within the uncertainty and an attenuated amplitude from base temperature up to 1.5 K (figs. S7 and S8). The temperature-dependent amplitudes of normalized oscillatory torque (Fig. 3A) and oscillatory MR (Fig. 3, B and C) are fitted by using the conventional LK formula (31). The fittings are reasonably good down to 60 mK, indicating that the LK expression, based on the Fermi liquid framework, appears to be valid in the Kondo insulator YbB$_{12}$. The agreement with the LK description confirms that the features we resolve are quantum oscillations rather than successive field-induced Lifshitz transitions.

The SdH oscillations are much more suppressed at higher temperatures as compared with dHvA oscillations at the same angle (Fig. 3, A and B), revealing a heavier effective mass in the electrical transport channel. The effective masses of the quasiparticles estimated from the dHvA and SdH

oscillations at the same tilt angle are ~6.6 m_e and 14.6 m_e, respectively. Therefore, it is unlikely that both types of oscillations originate from the same band. The SdH and dHvA frequencies have different angle dependences (Fig. 4). The dHvA frequencies F can be tracked with a two-dimensional (2D) Fermi surface model ($F \propto \frac{1}{\cos\theta}$) with the in-plane cross section area $A_{[001]} = 6.67$ nm^{-2} (Fig. 4, solid line). Given the lack of dHvA oscillations observed above $\simeq 20°$ (fig. S9), this inverse sinusoidal dependence can be explained by either a 2D Fermi cylinder or a heavily elongated 3D Fermi pocket. On the other hand, the angular dependence of the SdH frequencies displays a clearly nonmonotonic behavior: A frequency maximum appears at $\theta \sim 15°$ to 20° from the crystal axes, resulting in an "M"-shape with a local dip at $H \parallel [100]$ and a fast decrease in the frequency beyond the maximum. The 2D Fermi surface model apparently cannot describe this behavior. Our attempt to model the SdH frequencies (Fig. 4, dashed lines, and figs. S1 and S10) points to hyperbolic "neck" orbits [(21), section 8].

The amplitude of the quantum oscillations is determined by a combination of temperature, band curvature, Dingle, and spin factors. Assuming Dingle and spin factors do not depend on the band, the oscillation amplitude of the light band with the 2D character should be much larger than that of the heavy band with the 3D character. Because the observed SdH oscillations arise from

a relatively heavy band ($m^* \approx 15$ m_e), it is highly unusual that the oscillations observed in the dHvA effect, which detects the orbits from a much lighter band ($m^* \approx 6.6$ m_e), are not present in the SdH effect. We discuss in section 11 of (21) the possible origins of the dHvA oscillations, based on either surface state (32), charge-neutral Fermi surface (10), or a minority phase. Further, on the basis of the symmetry analysis and other tests in a even stronger magnetic field, the bulk SdH oscillations do not arise from metallic impurity phases [(21), sections 7, 9, and 10] or from a minority portion of sample, which has a lower I-M transition field [(21), sections 4 and 5].

There are several notable features in the SdH oscillations that we observed. The effective masses obtained from LK fittings are large, which is in agreement with the nature of a Kondo insulator in which strong electron correlations make the quasiparticles heavy. Even in the insulating state, a finite electronic specific heat coefficient γ is observed in YbB$_{12}$ (33). Assuming a spherical Fermi surface with $k_F \sim 0.156$ Å^{-1} ($F = 800$ T) and effective mass $m^* = 15$ m_e obtained from the SdH oscillations, the value of γ is calculated to be 7.6 mJ/mol K^2, comparable with the observed value of $\gamma \sim 8$ mJ/mol K^2 at 39 T (33). The background resistivity ρ still has a magnitude of more than 100 milliohm · cm above 40 T, which is well beyond that of normal metals (34). If we estimate the mean free path ℓ by considering a spherical

Fig. 1. Electrical transport and magnetic torque measurements in YbB$_{12}$. (**A**) Sketch of the experimental setup (21) and the definition of tilt angle ϕ with respect to the magnetic field. (Inset) Photograph of a YbB$_{12}$ single crystal (sample N2) mounted onto a cantilever beam magnetometer, with four gold wires attached to the crystallographic (001) surface for the transport measurement. (**B**) Resistivity of two YbB$_{12}$ single crystals plotted as a function of temperature. (Inset) The Arrhenius plot, lnρ versus $1/T$. According to the thermal activation model, the slope of the Arrhenius plot equals $\Delta/2k_B$, where Δ is the bandgap width and k_B is the Boltzmann constant. Linear fitting in two different temperature ranges, $20 < T < 40$ K and $6 < T < 2.5$ K, yields two characteristic gap widths, 12.5 and 4.7 meV, respectively. (**C**) Resistivity of YbB$_{12}$ sample N2 under different magnetic fields from 0 to 45 T, plotted against temperature. Open and solid symbols are data taken in ^{3}He cryostat at ϕ = 7.4° and in dilution fridge at ϕ = 8.5°, respectively. Solid lines are guides to the eye. (**D**) Magnetic torque in YbB$_{12}$ measured at $T = 350$ mK and at two different tile angles, ϕ = −6.8° and ϕ = 11.3°. Both exhibit strong quantum oscillations under high magnetic fields. The amplitude of the oscillatory part of magnetic torque at ϕ = 11.3° is ~6 × 10^{-8} N · m at the highest field, corresponding to an effective transverse magnetization of ~1.4 × 10^{-9} A · m^2 (1.51 × 10^{14} μ$_B$). (Inset) FFT on the magnetic torque signal with ϕ = 11.3° reveals a major peak at $F = 720$ T and its harmonics.
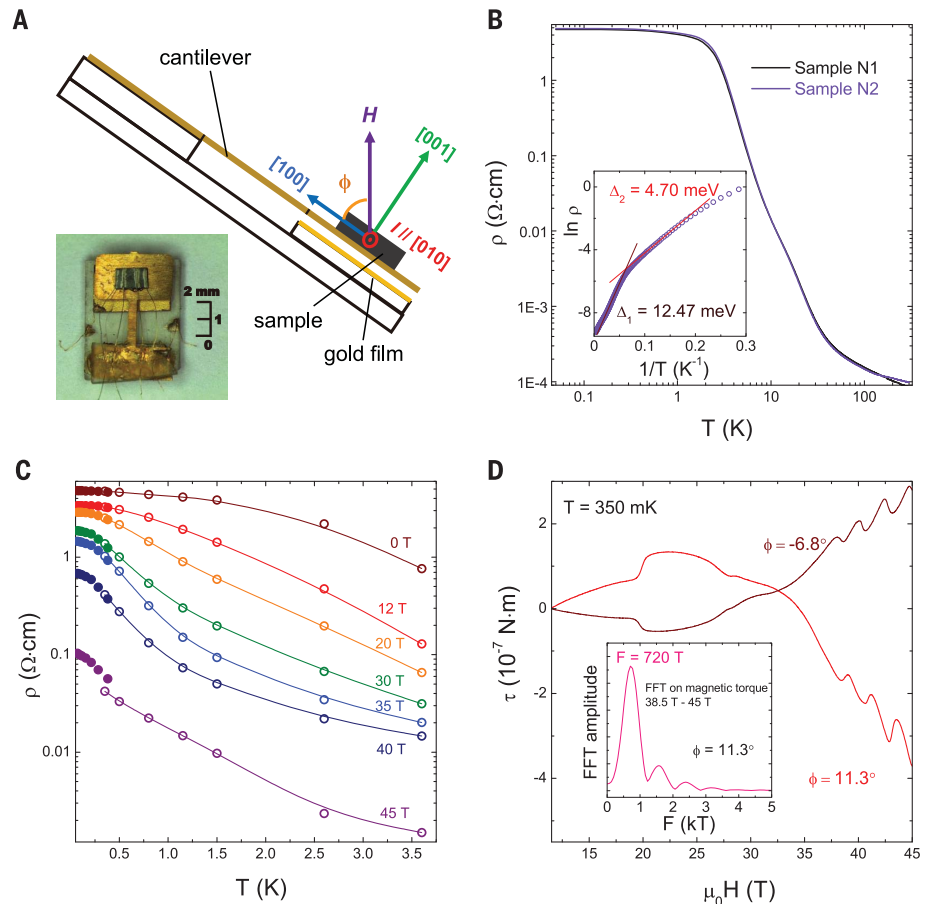
Fig. 2. Resistivity and electrical oscillations in intense magnetic fields in YbB₁₂. (**A**) Resistivity of sample N2 as a function of magnetic field measured up to 45 T taken at T = 350 mK at a tilt angle ϕ = 27.7°. Quantum oscillations are clearly observed at high magnetic field beyond 40.8 T. The extrema are marked by arrows. (Top inset) First magnetic field derivative of resistance has two prominent valleys and one peak. (Bottom inset) FFT on the magnetoresistance data presented between $\mu_0 H$ = 40.5 T and 45 T. A single peak frequency of F = 913 T is resolved. (**B**) Field derivative of sample resistivity at four different tilt angles. Dotted lines mark the approximately evenly spaced valleys of SdH oscillation. Three to four periods in total can be observed, depending on the field orientation. The oscillation pattern at ϕ is repeated at $90° − \phi$, which is consistent with the cubic symmetry of crystal structure.
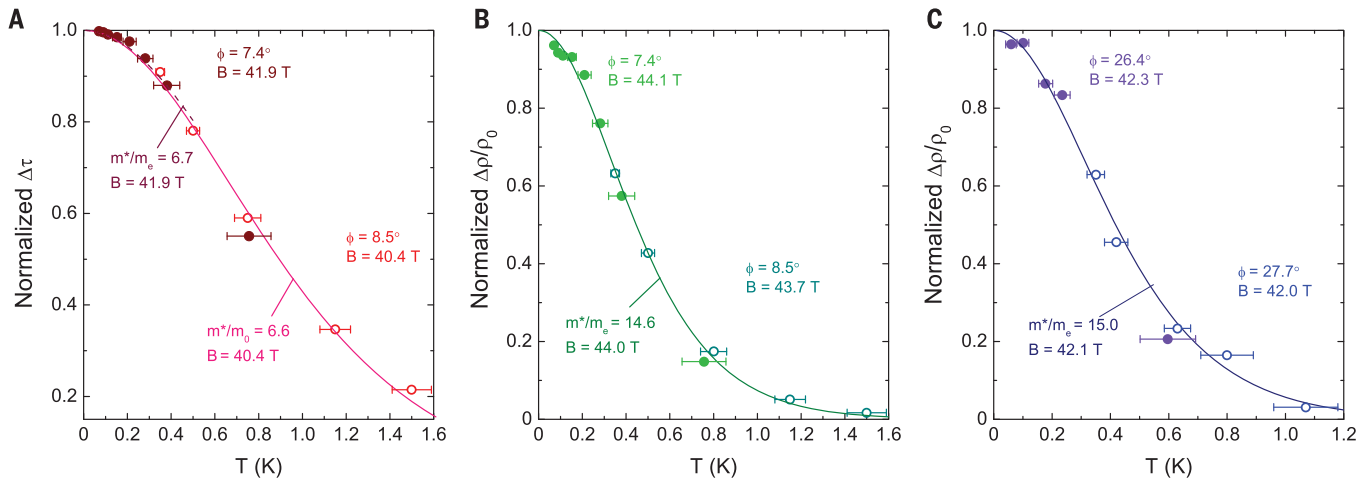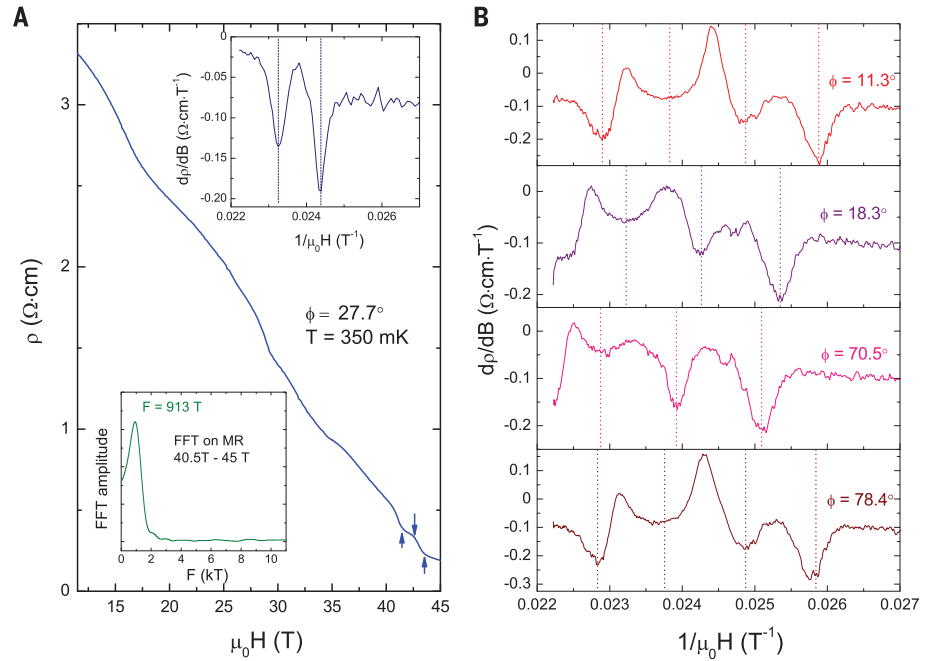
Fig. 3. Effective mass fitting on dHvA and SdH oscillations in YbB₁₂. (**A**) Normalized dHvA oscillation amplitude $\Delta\tau$ as a function of temperature. Open circles and solid circles are the data taken in ³He cryostat at ϕ = 8.5° and in portable dilution fridge at ϕ = 7.4°, respectively. The LK model fitting by using parameters m^* = 6.6 m_e, B = 40.4 T is indicated with a solid line based on the ³He cryostat data. In comparison, the LK fitting based on the portable dilution fridge data gives m^* = 6.7 m_e, B = 41.9 T, which is indicated with the dashed line. (**B** and **C**) Temperature dependence of normalized SdH amplitude $\Delta\rho/\rho_0$. (B) is taken at the same tilt angle as in (A), and (C) is taken at a tilt angle close to that in Fig. 2A. Open and solid symbols are data measured in portable ³He cryostat and dilution fridge, respectively. Here, ρ_0 is the zero-field resistivity at

corresponding temperatures. Solid lines are fittings based on the LK formula, with parameters shown in each panel. According to the fittings, SdH effective mass exhibits a weak anisotropy between $\phi \simeq 8°$ [(B), m^* = 14.6 m_e] and $\phi \simeq 27°$ and [(C), m^* = 15.0 m_e]. In (A) to (C), the quantum oscillation amplitudes (raw data are shown in fig. S8) are taken as the averaged value of an adjacent peak and valley obtained after subtracting a polynomial background from the raw data, and the effective magnetic field for each data set is an inverse average of the peak and valley positions. [Details of the background subtraction and the determination of the B parameters in the LK fittings are presented in (21), section 6.] The error bars on the temperature are estimated based on the MR effect on the ruthenium oxide thermometer above 11.4 T.

Fermi surface with m^* = 15 m_e and setting ρ = 0.4 ohm · cm (Fig. 2A), we will obtain an unphysically short mean free path, $\ell \sim 0.01$ nm.

Unconventional quantum oscillations have been proposed in insulators with hybridization gaps (6–11, 35, 36). However, our discovery of quantum oscillations in charge transport cannot be effectively interpreted by the theories that either

invoke charge-neutral quasiparticles (7–9) or associate the oscillations with the grand canonical potential (6). All these theories of exotic quantum oscillations predict that in a gapped system, the temperature dependence of the oscillation amplitude deviates from the conventional LK-formula at certain elevated temperatures. By contrast, our observations show that in YbB₁₂, the LK formu-

la works from T = 1.5 K (0.13 meV) down to T = 60 mK (0.0052 meV), a range of energy lower than both the hybridization gap width and the cyclotron energy of charge carriers at a field of $\simeq 40$ T (21), section 11]. Recent theories suggest that an emergent neutral Fermi sea can exist in a mixed-valence gapped system and exhibit both dHvA and SdH oscillations (10, 11), which may
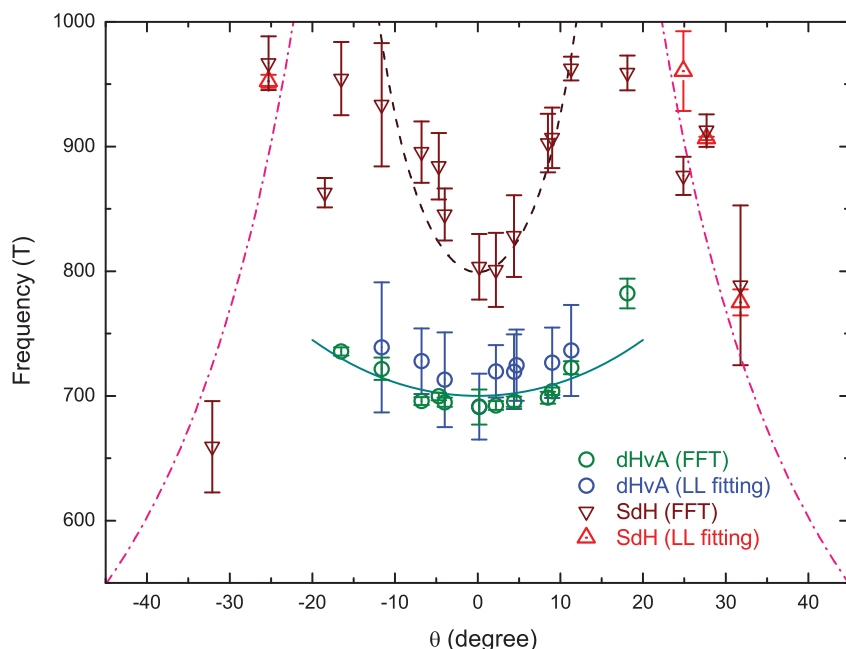
Fig. 4. Angular dependence of quantum oscillation frequencies. The frequencies of the quantum oscillations appearing at high field in YbB$_{12}$. Magnetic field H is rotated in a plane perpendicular to the current direction, and the effective tilt angle θ is defined as the angle between H and the equivalent crystal axes [001]/[100] in a cubic structure. Circles are dHvA frequencies obtained from FFT (green) and the slope of a linear fitting of the LL index versus inverse magnetic field (blue). The solid line is a calculation by using 2D Fermi surface model: $F = F_0/\cos\theta$ with $F_0 = 700$ T. Up triangles (red) and down triangles (brown) are SdH frequencies acquired from the FFT and from the linear fitting of the LL index plot, respectively. The dashed line is the fitting by a hyperboloid model representing a Fermi surface neck region with the principal axis along the [001] direction, whereas the dash-dot lines are simulation for the high-angle data points by using an oblate spheroid model with the principal axis along [001] direction. The detailed parameters in these models are provided in (21), section 8. The error bars come from the difference between different sampling windows for the FFT results and from the linear fitting error for the LL analysis.

shed light on the exotic SdH oscillations we resolved, although a deviation from LK formula is still required in this scenario. A proper theory is yet to be established to describe the quantum oscillations observed in YbB$_{12}$ under a high magnetic field.

REFERENCES AND NOTES

1. H. Tsunetsugu, M. Sigrist, K. Ueda, *Rev. Mod. Phys.* **69**, 809–864 (1997).
2. P. S. Riseborough, *Adv. Phys.* **49**, 257–320 (2000).
3. G. Li *et al.*, *Science* **346**, 1208–1212 (2014).
4. B. S. Tan *et al.*, *Science* **349**, 287–290 (2015).
5. Y. Luo, H. Chen, J. Dai, Z. Xu, J. D. Thompson, *Phys. Rev. B* **91**, 075130 (2015).
6. J. Knolle, N. R. Cooper, *Phys. Rev. Lett.* **115**, 146401 (2015).
7. J. Knolle, N. R. Cooper, *Phys. Rev. Lett.* **118**, 096604 (2017).
8. G. Baskaran, Majorana arXiv:1507.03477 [cond-mat.str-el] (2015).
9. O. Erten, P.-Y. Chang, P. Coleman, A. M. Tsvelik, *Phys. Rev. Lett.* **119**, 057603 (2017).
10. D. Chowdhury, I. Sodemann, T. Senthil, *Nat. Commun.* **9**, 1766 (2018).
11. I. Sodemann, D. Chowdhury, T. Senthil, *Phys. Rev. B* **97**, 045152 (2018).
12. M. Kasaya, F. Iga, K. Negishi, S. Nakai, T. Kasuya, *J. Magn. Magn. Mater.* **31-34**, 437–438 (1983).
13. M. Kasaya, F. Iga, M. Takigawa, T. Kasuya, *J. Magn. Magn. Mater.* **47-48**, 429–435 (1985).
14. T. Kasuya, *Europhys. Lett.* **26**, 277–281 (1994).
15. J.-M. Mignot *et al.*, *Phys. Rev. Lett.* **94**, 247204 (2005).
16. M. Okawa *et al.*, *Phys. Rev. B* **92**, 161108 (2015).
17. T. T. Terashima *et al.*, *J. Phys. Soc. Jpn.* **86**, 054710 (2017).
18. J. Yamaguchi *et al.*, *Phys. Rev. B* **79**, 125121 (2009).
19. Y. Utsumi *et al.*, *Phys. Rev. B* **96**, 155130 (2017).
20. H. Weng, J. Zhao, Z. Wang, Z. Fang, X. Dai, *Phys. Rev. Lett.* **112**, 016403 (2014).
21. Materials, methods, and additional data are available as supplementary materials.
22. J. C. Cooley, M. C. Aronson, Z. Fisk, P. C. Canfield, *Phys. Rev. Lett.* **74**, 1629–1632 (1995).
23. F. Iga, N. Shimizu, T. Takabatake, *J. Magn. Magn. Mater.* **177-181**, 337–338 (1998).
24. K. S. D. Beach, P. A. Lee, P. Monthoux, *Phys. Rev. Lett.* **92**, 026401 (2004).
25. K. Sugiyama, F. Iga, M. Kasaya, T. Kasuya, M. Date, *J. Phys. Soc. Jpn.* **57**, 3946–3953 (1988).
26. T. Takabatake *et al.*, *Phys. Rev. B Condens. Matter* **45**, 5740–5743 (1992).
27. G. S. Boebinger, A. Passner, P. C. Canfield, Z. Fisk, *Physica B* **211**, 227–229 (1995).
28. J. C. Cooley *et al.*, *Phys. Rev. B Condens. Matter* **52**, 7322–7327 (1995).
29. J. C. Cooley *et al.*, *J. Supercond.* **12**, 171–173 (1999).
30. S. Wolgast *et al.*, *Phys. Rev. B* **92**, 115110 (2015).
31. D. Shoenberg, *Magnetic Oscillations in Metals* (Cambridge Univ. Press, 2009).
32. K. Hagiwara *et al.*, *Nat. Commun.* **7**, 12690 (2016).
33. T. Terashima *et al.*, *Phys. Rev. Lett.* **120**, 257206 (2018).
34. M. Gurvitch, *Phys. Rev. B* **24**, 7404–7407 (1981).
35. L. Zhang, X.-Y. Song, F. Wang, *Phys. Rev. Lett.* **116**, 046404 (2016).
36. H. K. Pal, F. Piéchon, J.-N. Fuchs, M. Goerbig, G. Montambaux, *Phys. Rev. B* **94**, 125140 (2016).

PLASMONIC CHEMISTRY

Quantifying hot carrier and thermal contributions in plasmonic photocatalysis

Linan Zhou[1], Dayne F. Swearer[1], Chao Zhang[2], Hossein Robatjazi[2], Hangqi Zhao[2],
Luke Henderson[1], Liangliang Dong[1], Phillip Christopher[3], Emily A. Carter[4],
Peter Nordlander[2,5]*, Naomi J. Halas[1,2,5]*

Photocatalysis based on optically active, "plasmonic" metal nanoparticles has emerged as a promising approach to facilitate light-driven chemical conversions under far milder conditions than thermal catalysis. However, an understanding of the relation between thermal and electronic excitations has been lacking. We report the substantial light-induced reduction of the thermal activation barrier for ammonia decomposition on a plasmonic photocatalyst. We introduce the concept of a light-dependent activation barrier to account for the effect of light illumination on electronic and thermal excitations in a single unified picture. This framework provides insight into the specific role of hot carriers in plasmon-mediated photochemistry, which is critically important for designing energy-efficient plasmonic photocatalysts.

Numerous recent demonstrations of hot carrier–driven photocatalysis by metal nanoparticles (NPs) that support collective electronic excitations, known as surface plasmons, have greatly stimulated current research activity (1). Hot carriers produced by nonradiative decay of localized surface plasmons can be more energetic than those generated by direct photoexcitation (2). In contrast to semiconductor photocatalysis, a supralinear intensity dependence of photocatalytic reaction rates has been observed in multiple plasmonic photocatalytic reactions, demonstrating higher quantum yields with increasing photon flux (3–5).

In plasmonic photocatalysis, hot carrier–mediated chemical conversion can be synergistic with thermal excitation. For example, for plasmon-mediated O_2 dissociation on Ag cubes, the photocatalytic reaction rate increased exponentially with temperature under constant white-light illumination (5). Recent studies have begun to address the challenge of distinguishing thermal from nonthermal effects (6, 7). The apparent activation barrier for thermal catalysis can be decreased by illumination of a plasmonic photocatalyst (8–10). However, characterization of activation energies under illumination conditions to examine the mechanism of hot carrier–induced activation barrier reduction has been lacking.

We studied the effects of optical illumination on the apparent activation barrier E_a for ammonia decomposition ($2NH_3 \rightarrow N_2 + 3H_2$) using a plasmonic antenna-reactor (AR) photocatalyst (4). We varied the illumination wavelength λ and intensity I and measured the surface temperature T_s of the photocatalyst pellet with a thermal imaging camera to account for photothermal heating. Our analysis showed that photoinduced reductions of the reaction barrier have an electronic origin. The reaction mechanism of NH_3 decomposition is well studied, with no side reactions (11, 12), and NH_3 is a promising medium for CO_x-free hydrogen storage (13). With conventional thermal catalysis, this application faces substantial challenges because of the high E_a of NH_3 decomposition (1 to 2 eV), which requires high temperatures to achieve a turnover frequency (TOF) of 1 s^{-1} (one NH_3 molecule per active metal site per second) (14, 15).

Our plasmonic AR photocatalyst was a Cu–Ru surface alloy that consisted of a Cu NP antenna and Ru reactor sites (Cu-Ru-AR) (Fig. 1A), synthesized by coprecipitation (16) with subsequent annealing and reduction in the photoreaction chamber (supplementary materials). The photocatalytic reaction rate on Cu-Ru-AR was ~20 and ~177 times higher, respectively, relative to pure Cu and Ru NPs (Fig. 1B). For illumination at 9.6 W cm^{-2}, without external heating, the photocatalytic reaction rate of NH_3 decomposition over Cu-Ru-AR was as high as 1200 μmol H_2 g^{-1} s^{-1}. The reaction rate dropped within the noise of the measurement when the light was turned off. This process was repeated reproducibly multiple (five) times (Fig. 1C). The TOF based on Ru loading was >15 s^{-1}, and the energy efficiency and quantum yield were calculated to be 18 and 33.5%, respectively, under this set of conditions (see supplementary materials for calculation method). The ratio of photocatalytic reaction rates based on the measured amounts of NH_3, N_2, and H_2 were consistent with the stoichiometry of the reaction, confirming the absence of unintended side reactions.

To differentiate the contributions of plasmon-induced hot electrons and photothermal heating, the steady-state temperature on the photocatalyst surface was measured in situ with a thermal imaging camera (supplementary materials). T_s increased with I and reached 475.4°C at a peak intensity of 9.6 W cm^{-2} (fig. S11A). When NH_3 decomposition was performed without illumination, but with external heating temperatures equivalent to those achieved under illumination, the thermocatalytic rates of H_2 production were one to two orders of magnitude below the observed photocatalytic rates (Fig. 1D). Because of limited light penetration into the catalyst pellet (17), the effective amount of catalyst was appreciably smaller for photocatalysis than for thermocatalysis (in the dark). On the basis of this observation, we conclude that plasmon-induced hot carriers are the predominant effect that catalyzes NH_3 decomposition. Although the transient local temperature on the NP surface under illumination can be higher than the static temperature (18), the local temperature was calculated to be less than 0.1 K higher than that of the environment because the energy intensity of each laser pulse was very low (~0.12 μJ cm^{-2}) and essentially independent of NP size in this (sub–10-nm-diameter) regime (fig. S12, B to D) (19, 20).

To quantify the effect of illumination on E_a, we measured reaction rates over a range of surface temperatures by varying chamber temperature for each λ and I. We derived $E_a(\lambda, I)$ from the measured reaction rates and T_s values using the Arrhenius equation (table S2). Figure 2, A and B, shows the wavelength dependence of E_a for a fixed I at 3.2 W cm^{-2} and the intensity dependence of E_a for a fixed excitation wavelength at 550 nm, corresponding to the localized surface plasmon resonance (LSPR) frequency of the Cu-Ru-AR photocatalyst (λ = 550 nm, fig. S4). Without light, the E_a was 1.21 eV (black line in Fig. 2A). Illumination on resonance at an intensity of 3.2 W cm^{-2} led to the greatest reduction of E_a, from 1.21 to 0.35 eV. For longer λ, the reduction in E_a is less because of reduced optical absorption that decreased hot carrier generation. At shorter λ, the absorption was still high, owing to Cu interband transitions (21), but the reduction in E_a was smaller, because the energetic electrons produced by interband transitions in Cu have substantially lower energies than those produced by plasmon decay (2). This is in contrast to recent observations that more, but less energetic, carriers generated by interband transitions in Au NPs were more efficient than those generated by plasmon resonant excitation in solution phase catalysis (22). In each case, however, the specific energetics of the reaction itself are likely to be very important in ultimately determining photocatalytic efficiencies. For resonant illumination (Fig. 2B), E_a decreased with increasing light intensity. A three-dimensional (3D) contour map of the light-induced E_a (Fig. 2C) shows that the wavelength dependence is similar for all light intensities and that the trend of decreasing E_a with increasing

[1]Department of Chemistry, Rice University, Houston, TX 77005, USA. [2]Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005, USA. [3]Department of Chemical Engineering, University of California, Santa Barbara, Santa Barbara, CA, 93106-5080, USA. [4]School of Engineering and Applied Science, Princeton University, Princeton, NJ 08544-5263, USA. [5]Department of Physics and Astronomy, Rice University, Houston, TX 77005, USA.
*Corresponding author. Email: nordland@rice.edu (P.N.); halas@rice.edu (N.J.H.)

Fig. 1. Catalytic ammonia decomposition.
(**A**) Schematic of the structure of Cu-Ru-AR consisting of a Cu NP antenna with a Cu-Ru surface alloy, where the Ru provides the reactor sites. (**B**) H_2 formation rate of photocatalysis (9.6 W cm^{-2}) and thermocatalysis (482°C) on Cu-Ru-AR, Cu, and Ru NPs. (**C**) Multiple-hour measurement of photocatalytic rates on Cu-Ru-AR under 9.6 W cm^{-2} white-light illumination without external heating. (**D**) Comparison of photocatalytic and thermocatalytic rates on Cu-Ru-AR. The horizontal axis corresponds to the surface temperature of the catalyst caused by photothermal heating (photocatalysis) or external heating (thermocatalysis). The light intensity differences between successive data points are 0.8 W cm^{-2}.
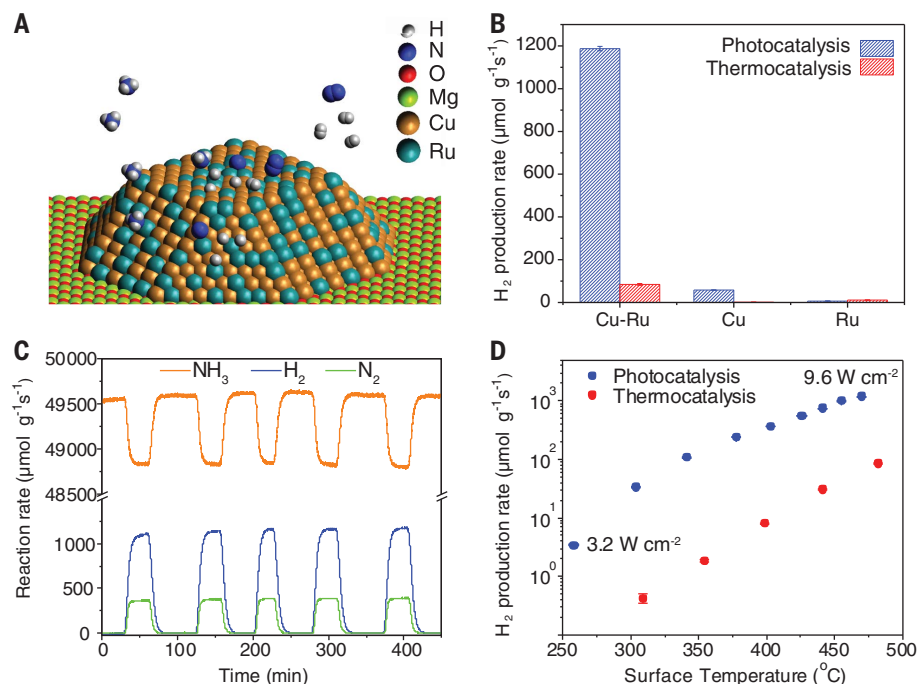
Fig. 2. Light-dependent activation barrier.
(**A** and **B**) Arrhenius plots of apparent activation barriers for (A) different wavelengths under constant intensity of 3.2 W cm^{-2} and (B) various light intensities at 550 nm. The black line represents thermocatalysis in the dark. The surface temperatures of the catalyst were measured with an infrared camera and used for the Arrhenius analysis. (**C**) A 3D representation of $E_a(\lambda, I)$ for different wavelengths and intensities through interpolation of 46 data points. (**D**) Reaction order with respect to P_{NH_3} in photocatalysis (6.4 W cm^{-2} white light) and thermocatalysis (427°C). r, H_2 production rate; n, reaction order.
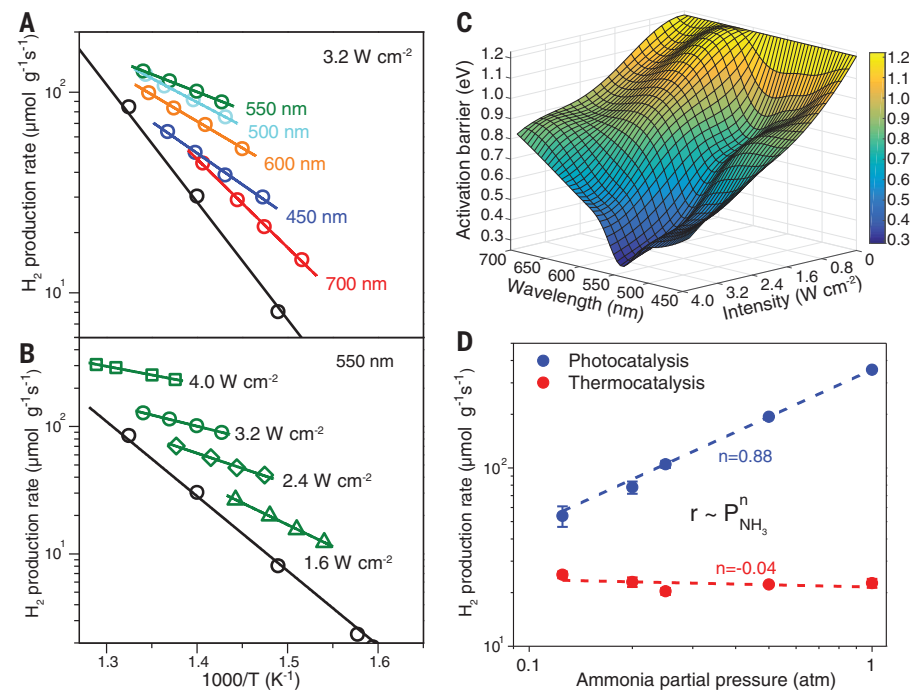
I holds for all λ studied. Under optimal illumination, at the LSPR frequency and an intensity of 4 W cm^{-2}, E_a was reduced to ~0.27 eV.

By mapping E_a as a function of illumination conditions (Fig. 2C), we could predict catalytic performance. For example, according to the Arrhenius curve obtained for 3.2 W cm^{-2} illumination at 550 nm (green line in Fig. 2A), the catalytic reaction rate was calculated to be 22.7 μmol H_2 g^{-1} s^{-1} at 293°C, which is also the T_s for this illumination condition without external heating (fig. S11C). This value was in good agreement with

the measured reaction rate of 19.0 μmol H_2 g^{-1} s^{-1} (fig. S16). Furthermore, because the light-induced heating of nanostructures for arbitrary illumination conditions can be calculated (23, 24), a knowledge of the light-induced $E_a(\lambda, I)$ can provide a direct relation between total energy input and reaction rates. With this insight, it is possible to optimize the efficiency of plasmonic photocatalysis in terms of overall energy efficiency (per mole product divided by per unit energy input).

In the present study, $E_a(\lambda, I)$ was determined by measuring the T_s of the catalyst during the

reaction, making it possible to clearly distinguish the hot carrier contributions from photothermal heating. In situations where the temperature change of the catalyst cannot be determined independently or through calculations, the measured light dependence of the activation barrier will also include contributions from photothermal heating. In the supplementary materials, we discuss how the intrinsic light-dependent barrier can be determined from the measured barrier in that case.

The E_a of a reaction is defined as the sum of the activation barriers for the rate-determining
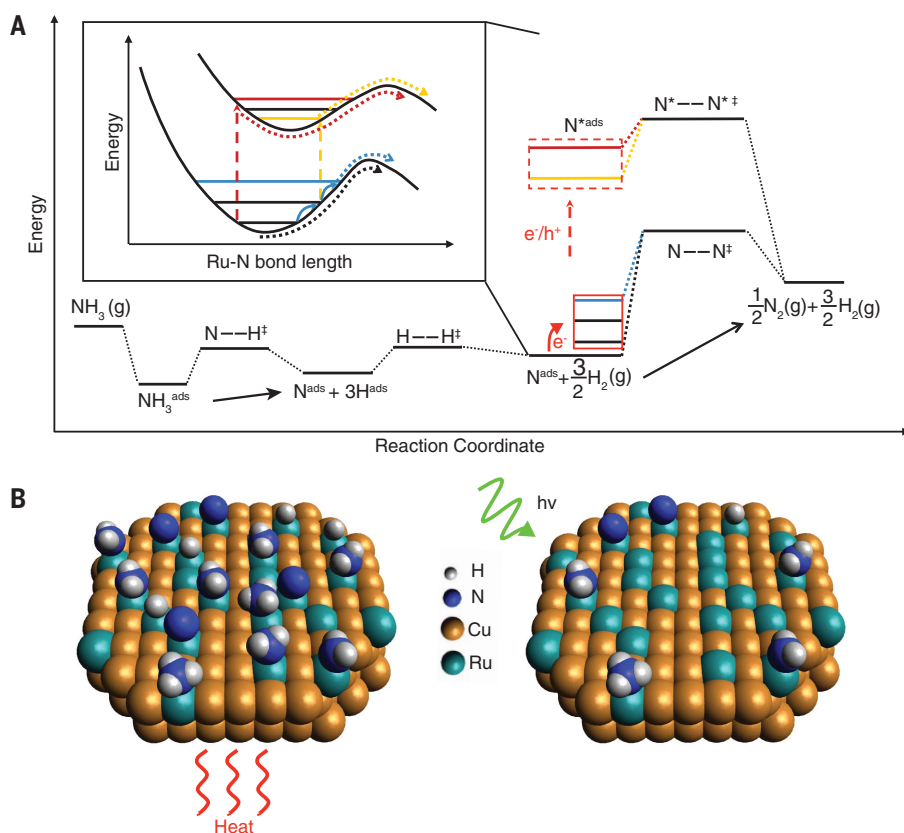
Fig. 3. Mechanisms for hot carrier–mediated reduction of activation barrier. (**A**) Schematic energetics of elementary reaction steps for NH_3 decomposition. The ‡ denotes transition state, and the two relevant RDSs, N–H bond scission and associative desorption of N_2, are labeled. N* refers to excited Ru–N surface species: either vibronic levels in an excited electronic state (red dashed arrow and box) or vibrational excitations in the electronic ground state (red solid arrow and box). The activation barriers after excitation (red, yellow, and blue dotted lines) are all lower than that of the ground state (black dotted line). The Franck-Condon diagram inset illustrates the mechanisms in more detail. Dark red and yellow dashed arrows denote possible hot carrier transfer into or electronic excitation of the Ru–N surface species. Blue solid arrows denote multiple vibrational excitations of the electronic ground state. e^-, electron; h^+, hole. (**B**) Schematics of surface coverage under heating versus light-illumination conditions. In contrast to thermocatalysis (left), photocatalysis and hot carrier generation (right) can promote desorption of "poisoning" adsorbed intermediates. hv, light.

step (RDS) ($E_{a\text{-RDS}}$), the enthalpy of the steps that produce species involved in the RDS, and a coverage-dependent term related to the enthalpy required to clean active sites occupied by reaction intermediates (*25*). For NH_3 decomposition on a Ru surface, there are two reported RDSs: (i) N–H bond scission (one of the three N–H breaking steps) and (ii) associative desorption of N_2 (*26, 27*). Measurements of the reaction rate dependence on NH_3 partial pressure (P_{NH_3}) show a zeroth order dependence for the thermocatalytic reaction and a first-order dependence for the photocatalytic reaction (Fig. 2D). In the thermocatalytic experiments, the RDS is the associative desorption of N_2, with the reaction intermediates, mostly N, adsorbed on the surface, blocking reaction sites (*28, 29*). This model was confirmed by the large measured E_a (1.21 eV), which stems from the sum of $E_{a\text{-RDS}}$ and the energy penalty associated with cleaning active sites occupied by adsorbed intermediates. For photocatalysis with 6.4 W cm^{-2} white-light illumination (T_s of 403°C, fig. S11B), the reaction order with respect to P_{NH_3} increased substantially, to 0.88. This increase in reaction order, along with a reduction in E_a upon illumination, suggests that hot carriers induced by resonant illumination modify the reaction kinetics by reducing the activation barrier for associative desorption of N_2 ($E_{a\text{-RDS}}$), while simultaneously decreasing the surface coverage of adsorbed intermediates. It is plausible that higher-intensity illumination shifted the RDS from associative desorption of N_2 to N–H bond scission. Both processes decrease E_a, syner-

gistically promoting efficient plasmon-mediated NH_3 decomposition.

Hot carriers can influence the energetics of adsorbed N (N^{ads}) and enhance the net reaction rate, as verified by an analysis of the microkinetics for this reaction (*30*) discussed in the supplementary materials. More importantly, the $E_{a\text{-RDS}}$, and thus the E_a, can be reduced by activating the N^{ads} through hot carriers, in many ways, along the reaction pathway (Fig. 3A). Hot electrons and holes can transfer into N^{ads} and electronically excite the Ru–N surface species, thus facilitating the N_2 desorption process (red dashed arrow in Fig. 3A). The Ru–N bond activation then occurs through an excited state with a lower activation barrier (dark red and yellow dotted lines in Fig. 3A). This process requires resonant conditions between the hot carriers and the acceptor levels of the Ru–N surface species (dark red and yellow dashed arrows in the inset). Once created by plasmon decay, hot carriers rapidly relax toward the Fermi level through a carrier multiplication process induced by electron-electron scattering (*31*). Electron-phonon scattering on a monometallic particle plays a relatively minor role, because such phonons involve atoms of the same charges and do not possess electrical multipolar moments. However, the interaction can be strong for localized vibrational modes involving atoms of different charges, such as the Ru–N surface species.

Thus, another possibility for activation of the Ru–N bond is inelastic electron-vibrational dipole scattering (red solid arrow in Fig. 3A). This pro-

cess does not require resonant conditions and can occur as long as the hot carriers have energies greater than the vibrational quantum. The excited electrons can induce multiple vibrational transitions of the Ru–N bond (blue solid arrows in the Franck-Condon diagram inset in Fig. 3A), and as the vibrational energy stored in the bond increases, the activation energy is reduced. In particular, a recent report showed that the excited vibrational state of surface species could exhibit 100-ps-long lifetimes and thus assist the barrier-crossing process (*32*). Overall, both excitations (hot carrier, energy transfer, and electron-vibrational scattering) could contribute to accelerating the associative desorption of N_2, which has been described previously as desorption induced by electronic transitions (DIET) on metal substrates under light excitation (*33–36*). With more molecules excited to higher vibronic states by higher light intensities, closer-to-resonant photoexcitation, or both processes, the $E_{a\text{-RDS}}$ decreases, as does the E_a of the entire reaction. Although the DIET mechanism would also apply to other species such as H_2, desorption of these species is not the RDS here and would have little effect on the E_a. Hot carrier–induced desorption of intermediates (Fig. 3B) also contributes to the decrease of E_a by decreasing the coverage of intermediates and releasing the active sites, as discussed in the supplementary materials.

REFERENCES AND NOTES

1. Y. Zhang *et al.*, *Chem. Rev.* **118**, 2927–2954 (2017).
2. B. Y. Zheng *et al.*, *Nat. Commun.* **6**, 7797 (2015).

3. C. Zhang et al., Nano Lett. **16**, 6677–6682 (2016).
4. D. F. Swearer et al., Proc. Natl. Acad. Sci. U.S.A. **113**, 8916–8920 (2016).
5. P. Christopher, H. Xin, A. Marimuthu, S. Linic, Nat. Mater. **11**, 1044–1050 (2012).
6. X. Zhang et al., Nano Lett. **18**, 1714–1723 (2018).
7. Y. Yu, V. Sundaresan, K. A. Willets, J. Phys. Chem. C **122**, 5040–5048 (2018).
8. X. Zhang et al., Nat. Commun. **8**, 14542 (2017).
9. H. Song et al., ACS Appl. Mater. Interfaces **10**, 408–416 (2018).
10. Y. Kim, D. Dumett Torres, P. K. Jain, Nano Lett. **16**, 3399–3407 (2016).
11. G. Ertl, M. Huber, J. Catal. **61**, 537–539 (1980).
12. M. C. J. Bradford, P. E. Fanning, M. A. Vannice, J. Catal. **172**, 479–484 (1997).
13. T. E. Bell, L. Torrente-Murciano, Top. Catal. **59**, 1438–1457 (2016).
14. F. Schüth, R. Palkovits, R. Schlögl, D. S. Su, Energy Environ. Sci. **5**, 6278–6289 (2012).
15. A. Amano, H. Taylor, J. Am. Chem. Soc. **76**, 4201–4204 (1954).
16. M. Behrens, Catal. Today **246**, 46–54 (2015).
17. H. Robatjazi et al., Nat. Commun. **8**, 27 (2017).
18. V. K. Pustovalov, Chem. Phys. **308**, 103–108 (2005).
19. A. M. Goodman et al., ACS Nano **11**, 171–179 (2017).
20. S. C. Nguyen et al., ACS Nano **10**, 2144–2151 (2016).
21. G. H. Chan, J. Zhao, E. M. Hicks, G. C. Schatz, R. P. Van Duyne, Nano Lett. **7**, 1947–1952 (2007).
22. J. Zhao et al., ACS Cent. Sci. **3**, 482–488 (2017).
23. H. I. Villafán-Vidales, S. Abanades, C. Caliot, H. Romero-Paredes, Appl. Therm. Eng. **31**, 3377–3386 (2011).
24. F. J. Valdés-Parada, H. Romero-Paredes, G. Espinosa-Paredes, Int. J. Hydrogen Energy **36**, 3354–3363 (2011).
25. H. Lynggaard, A. Andreasen, C. Stegelmann, P. Stoltze, Prog. Surf. Sci. **77**, 71–137 (2004).
26. W. Tsai, W. H. Weinberg, J. Phys. Chem. **91**, 5302–5307 (1987).
27. C. Egawa, T. Nishida, S. Naito, K. Tamaru, J. Chem. Soc., Faraday Trans. I **80**, 1595–1604 (1984).
28. A. Hellman et al., Surf. Sci. **603**, 1731–1739 (2009).
29. V. Prasad, A. M. Karim, A. Arya, D. G. Vlachos, Ind. Eng. Chem. Res. **48**, 5255–5265 (2009).
30. C. T. Campbell, ACS Catal. **7**, 2770–2779 (2017).
31. J. G. Liu, H. Zhang, S. Link, P. Nordlander, ACS Photonics **5**, 2584–2595 (2017).
32. P. R. Shirhatti et al., Nat. Chem. **10**, 592–598 (2018).
33. H. L. Dai, W. Ho, Laser Spectroscopy and Photochemistry on Metal Surfaces (World Scientific, 1995).
34. C. Frischkorn, M. Wolf, Chem. Rev. **106**, 4207–4233 (2006).
35. S. A. Buntin, L. J. Richter, R. R. Cavanagh, D. S. King, Phys. Rev. Lett. **61**, 1321–1324 (1988).
36. V. A. Spata, E. A. Carter, ACS Nano **12**, 3512–3522 (2018).

ADAPTIVE RADIATIONS

Rapid Pliocene adaptive radiation of modern kangaroos

Aidan M. C. Couzens[1,2]* and Gavin J. Prideaux[1]

Differentiating between ancient and younger, more rapidly evolved clades is important for determining paleoenvironmental drivers of diversification. Australia possesses many aridity-adapted lineages, the origins of which have been closely linked to late Miocene continental aridification. Using dental macrowear and molar crown height measurements, spanning the past 25 million years, we show that the most iconic Australian terrestrial mammals, "true" kangaroos (Macropodini), adaptively radiated in response to mid-Pliocene grassland expansion rather than Miocene aridity. In contrast, low-crowned, short-faced kangaroos radiated into predominantly browsing niches as the late Cenozoic became more arid, contradicting the view that this was an interval of global browser decline. Our results implicate warm-to-cool climatic oscillations as a trigger for adaptive radiation and refute arguments attributing Pleistocene megafaunal extinction to aridity-forced dietary change.

Adaptive radiation within emerging environments is believed to underpin the diversification of many animal groups (1). Radiating lineages can diverge rapidly (2), but this may be difficult to detect if phylogenetic analyses are poorly calibrated or models are misspecified (3, 4). Often the result is a dichotomy between ancient (short-fuse) and younger, more rapid (longer-fuse) diversification models, which can describe starkly different scenarios of ecologically driven radiation (4).

Over the past 15 million years, Australia has undergone a shift from a mesic to a largely arid continent (5). This environmental transition has been closely linked with the evolution of a diverse, aridity-adapted biota (6), which includes the kangaroos and wallabies (Macropodidae; Fig. 1A), the most abundant and diverse marsupial herbivores to evolve. But the timing and drivers of macropodid evolution have been difficult to resolve because of a patchy fossil record and imprecise divergence dating (7). Nonetheless, most phylogenies infer that the grass-feeding kangaroos (Macropodini), constituting more than one-third of macropodid diversity, underwent a pronounced late Miocene diversification (7–9), 3 million to 8 million years before Australian grasslands emerged (10).

In North America and Eurasia, Neogene grassland expansion ushered in the diversification of herbivores with high-crowned molar teeth, most notably ungulates (11, 12). Ungulate success has been linked to the improved resistance of their teeth to the elevated dental wear characteristic of grass-based diets (13). Similarly, extant grass-feeding kangaroos have higher-crowned molars (14) (Fig. 1B) and are much more diverse than their browsing counterparts. This suggests that grass exploitation has been a key factor in macropodin success (7).

To test the roles of late Miocene aridity and Pliocene grassland expansion in the adaptive radiation of kangaroos, we measured molar crown height and macrowear (figs. S1 and S2) for >1000 macropodoid (kangaroos and relatives) specimens from the modern fauna and 93 fossil assemblages (table S1) of late Oligocene to late Pleistocene age (15). Late Oligocene balbarines (plesiomorphic lophodont macropodoids) and basal macropodids have low-crowned dentitions and low macrowear levels (Fig. 2, B and C) indicative of low-abrasion diets. Crown height disparity increased in the early Miocene (Fig. 2B), probably as a response to more exploitation of fibrous plant resources, especially given that this coincides with the evolution of shearing bilophodont macropodid dentitions (16). Increased middle Miocene balbarine macrowear seems to capture an attempt by these plesiomorphic kangaroos to switch to more abrasive plant resources near the middle Miocene climatic optimum (Fig. 2, A and C), shortly before their extinction.

Unexpectedly, there is no late Miocene increase in crown height or macrowear (Fig. 2, B and C) analogous to those interpreted as responses to aridity among Northern Hemisphere herbivores (11, 17). Instead, late Miocene macropodids, represented by early sthenurines (short-faced kangaroos) and macropodines (modern kangaroos and wallabies), have even lower-crowned and less worn teeth than earlier macropodoids. This is indicative of browse- or even fungi-based diets (Fig. 1B) among central Australian faunas, which should have been among the first to experience grassland expansion (18).

Marked increases in crown height and macrowear across the Miocene-Pliocene transition (Fig. 2, B and C) herald a major adaptive shift in kangaroo evolution. Macropodine crown height increased by up to 40% in as little as 3 million years,
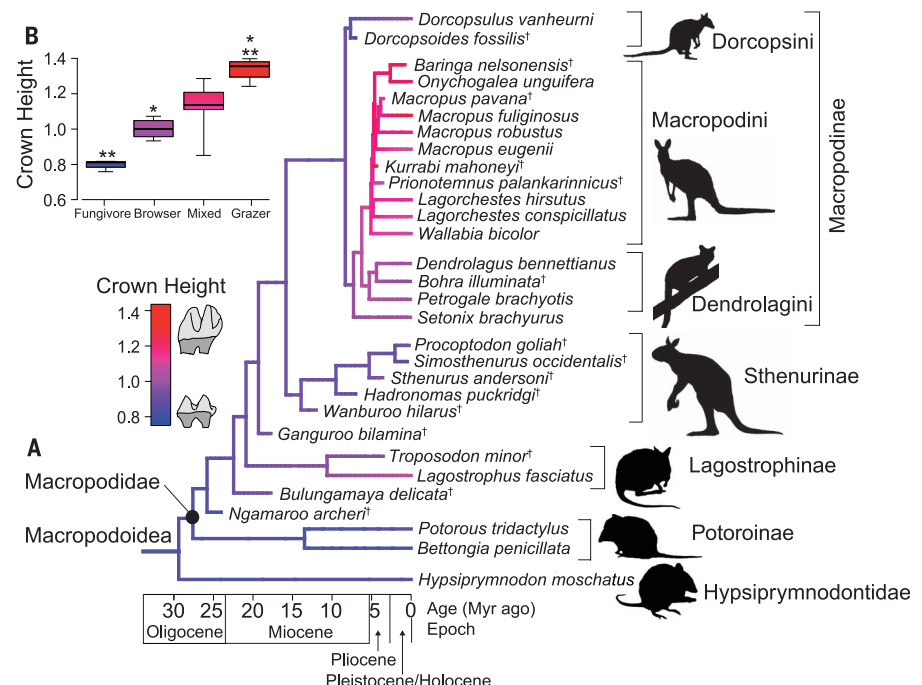
Fig. 1. Phylogenetic relationships and reconstructed crown height evolution in Macropodoidea. (**A**) Ancestral-state reconstruction of macropodoid molar crown height. Myr, millions of years. (**B**) Box plot of molar crown height and diet relationships among extant macropodoids. Diet categories with the same asterisk notation differ from each other at 2σ based on phylogenetic generalized least-squares regression (tables S5 and S6). Daggers denote extinct taxa. Colors in (A) match those of (B). See (15) for reference image attribution.

[1]College of Science and Engineering, Flinders University, Bedford Park, South Australia 5042, Australia. [2]Naturalis Biodiversity Center, 2300 RA Leiden, Netherlands.
*Corresponding author. Email: aidan.couzens@naturalis.nl
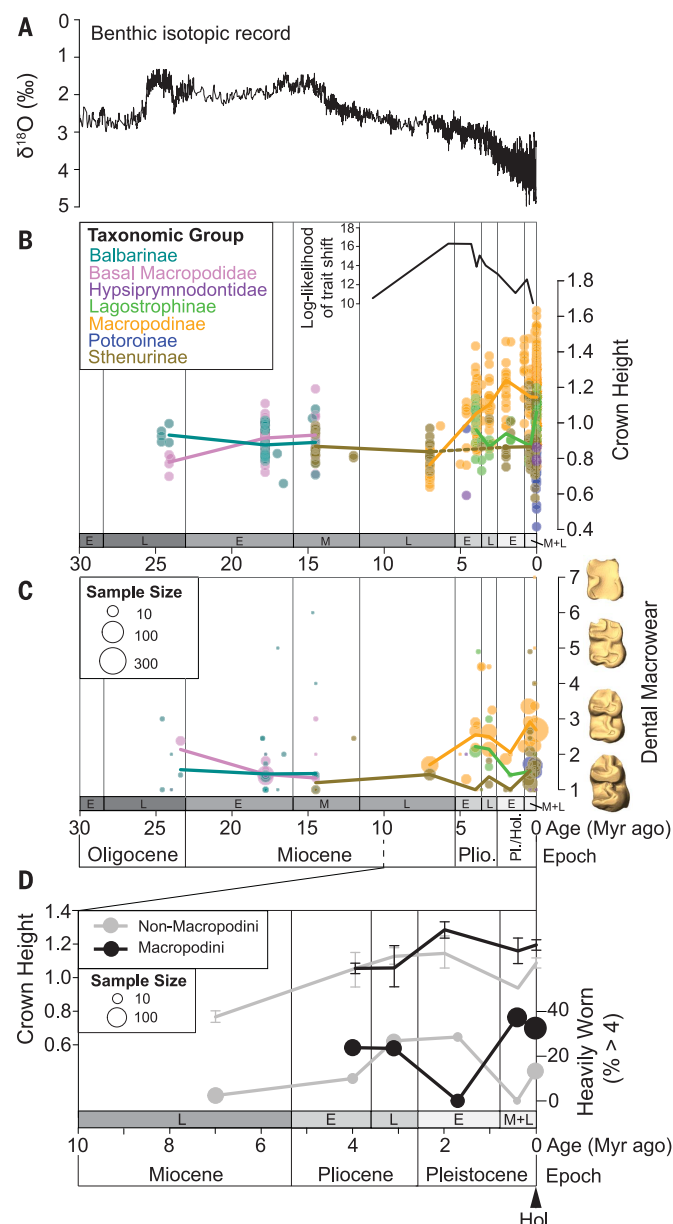
Fig. 2. Late Cenozoic macropodoid crown height and macrowear time series. (**A**) Benthic stable oxygen isotopic curve (δ^{18}O, black) (30). (**B**) Individual mandibular crown height measurements. (**C**) Geometric mean macrowear scores for each assemblage. Trend lines are clade averages for sub-epoch bins. (**D**) Late Neogene and Quaternary dental evolution within Macropodinae. E, M, and L denote early, middle, and late.

Fig. 3. Alternative models of kangaroo diversification and late Cenozoic climate. (**A**) Benthic stable oxygen isotopic record (30) and northwest Australian paleoenvironmental proxies (10); ky, 1000 years. (**B**) A "short-fuse" hypothesis linking macropodin divergence with late Miocene aridification. Divergence times follow (8). (**C**) An "explosive" hypothesis linking macropodin diversification with early Pliocene warming and mid-Pliocene grassland expansion. (**D**) Simulated macropodine crown height under "short-fuse" and "explosive" diversification models.

---

similar to rates measured among European and North American ungulates (12, 17). The Pliocene shift toward more heavily worn teeth is not explained by taphonomic biases (fig. S7) or the weak correlation with increased longevity (fig. S8). Instead, it closely tracks macrowear differences between extant browsing and grass-feeding kangaroos (fig. S9). Simulations that vary the position of dietary selection across the kangaroo phylogeny (fig. S6) point to an adaptive transition at the base of Macropodini as the driver of rapid crown height evolution (table S7). The associated

increases in dental wear and crown height strongly suggest that this adaptive transition was linked with intensified selection for dental durability. Increased crown height was probably favored because it delayed loph collapse, when bilophodont molars are relegated to a crushing rather than cutting modality (19). We infer that increased grass exploitation was the ecological driver because the most pronounced trait shifts (Fig. 2B) align with isotopic and pollen evidence for mid-Pliocene grassland expansion (10) (Fig. 3A). However, the relationship be-

tween diet and crown height (or macrowear) among extant kangaroos (Fig. 1B and fig. S9) implies that most Pliocene macropodines were probably consuming both grass and dicot leaves, consistent with enamel δ^{13}C values from the chinchilla assemblage (20). Not until the Pleistocene are macropodins differentiated from other macropodines by heavily worn, higher-crowned teeth (Fig. 2D), which suggests a much more recent origin of specialized, grazing diets among kangaroos than among Northern Hemisphere herbivores such as horses (12). The link

between generalist Pliocene kangaroo diets and rapid dental evolution highlights how opportunistically consumed foods can be potent drivers of dietary adaptation (21).

Most kangaroo phylogenies fit a "short-fuse" radiation model (Fig. 3B), where macropodin generic splits, and many intrageneric splits, occur in the drying late Miocene (7–9). However, this fits poorly with evidence for a Pliocene adaptive shift and the absence of any known Miocene macropodins. Our data instead suggest that most modern macropodin genera emerged during the early Pliocene warm phase (22) and then adaptively radiated during the arid late Pliocene and early Pleistocene as C_4-photosynthesizing grasses expanded (10) (Fig. 3, A and C). This scenario implies a rapid origination of the modern macropodin genera (*Lagorchestes, Macropus, Onychogalea, Setonix, Wallabia*) within as little as a million years, with subsequent mid-Pliocene grassland expansion providing ecological opportunity for macropodin cladogenesis and hybridization (23).

The limited extent of the Australian late Miocene fossil record means that a hitherto concealed pre-Pliocene macropodin radiation cannot be ruled out, but several lines of evidence favor a Pliocene "explosive" diversification model. First, despite >50 years of collecting, the late Miocene Alcoota assemblage of central Australia has yielded no macropodins but does include many specimens of the low-crowned dorcopsin *Dorcopsoides fossilis*, sthenurine *Hadronomas puckridgi*, and three as yet undescribed, low-crowned, non-macropodin kangaroos. Second, a rapid Pliocene radiation would help to explain the persistent phylogenetic lability of *Setonix, Onychogalea*, and *Wallabia*. Third, a rapid Pliocene diversification would require fewer independent convergences on grass feeding, versus at least six required under a "short-fuse" scenario (fig. S10). Finally, trait simulations under a "short-fuse" hypothesis predict high-crowned macropodins well in advance of paleo-environmental evidence for grassland expansion, whereas an "explosive" model much more closely tracks crown height dynamics observed in the fossil record (Fig. 3D).

The late Cenozoic has been interpreted as a phase of waning browser diversity (11), but we find that the archetypal browsing kangaroos, the Sthenurinae, were adaptively radiating through this interval. Macrowear and crown height data

indicate that middle and late Pleistocene sthenurines were consuming more abrasive plants than earlier in the late Cenozoic, although low-abrasion browse remained their staple diet (Fig. 2, B and C). This Pleistocene dietary expansion aligns with a doubling of sthenurine species richness (24) at a time when Australia was becoming increasingly arid (5, 25). The association of dietary change with dental adaptation and increasing species richness, amid deepening Pleistocene aridity, discounts the supposed importance of aridity-forced dietary change in sthenurine extinction (26). Microwear, isotopic, and morphological evidence point to reliance of some middle and late Pleistocene sthenurine kangaroos on chenopod shrubs such as saltbush, which are adapted to low-rainfall, high-salinity conditions (24, 27). We speculate that a Pleistocene expansion of chenopod biomass may explain why, unlike in North America (11), Australian browser diversity increased in spite of declining ecosystem productivity.

Aridity has been widely implicated in recent terrestrial diversifications (6, 28), but our data reveal a more dynamic picture, where warm-to-cool oscillations promote taxonomic diversification followed by ecological and morphological diversification. Warm-wet intervals may prime clades for rapid adaptation during ensuing arid intervals, perhaps by fostering trophic generalists that can radiate when ecological opportunity arrives (29). Tests of this model, which leverage the Cenozoic record of oscillating climate, hold promise for uncovering how climatically driven ecological change promotes adaptive radiation.

REFERENCES AND NOTES

1. D. Schluter, *The Ecology of Adaptive Radiation* (Oxford Univ. Press, 2000).
2. S. Gavrilets, J. B. Losos, *Science* **323**, 732–737 (2009).
3. R. M. D. Beck, M. S. Y. Lee, *Proc. R. Soc. B* **281**, 20141278 (2014).
4. M. J. Phillips, *Syst. Biol.* **65**, 546–557 (2016).
5. H. A. Martin, *J. Arid Environ.* **66**, 533–563 (2006).
6. M. Byrne *et al.*, *Mol. Ecol.* **17**, 4398–4417 (2008).
7. G. J. Prideaux, N. M. Warburton, *Zool. J. Linn. Soc.* **159**, 954–987 (2010).
8. R. W. Meredith, M. Westerman, M. S. Springer, *Aust. J. Zool.* **56**, 395–410 (2009).
9. W. G. Dodt, S. Gallus, M. J. Phillips, M. A. Nilsson, *Sci. Rep.* **7**, 16811 (2017).
10. J. W. Andrae *et al.*, *Geophys. Res. Lett.* **45**, 4831–4840 (2018).
11. C. M. Janis, J. Damuth, J. M. Theodor, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 7899–7904 (2000).
12. M. C. Mihlbachler, F. Rivals, N. Solounias, G. M. Semprebon, *Science* **331**, 1178–1181 (2011).
13. C. M. Janis, M. Fortelius, *Biol. Rev. Camb. Philos. Soc.* **63**, 197–230 (1988).
14. C. M. Janis, *Aust. Mammal.* **13**, 49–53 (1990).
15. See supplementary materials.
16. A. M. C. Couzens, A. R. Evans, M. M. Skinner, G. J. Prideaux, *Evolution* **70**, 568–585 (2016).
17. J. Jernvall, M. Fortelius, *Nature* **417**, 538–540 (2002).
18. N. Herold, M. Huber, D. R. Greenwood, R. D. Müller, M. Seton, *Geology* **39**, 3–6 (2011).
19. W. von Koenigswald, *Ann. Zool. Fenn.* **51**, 162–176 (2014).
20. S. Montanari, J. Louys, G. J. Price, *PLOS ONE* **8**, e66221 (2013).
21. P. S. Ungar, in *Comparative Dental Morphology*, T. Koppe, G. Meyer, K. W. Alt, Eds. (Karger, 2009), pp. 38–43.
22. J. M. K. Sniderman *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 1999–2004 (2016).
23. M. A. Nilsson, Y. Zheng, V. Kumar, M. J. Phillips, A. Janke, *Genome Biol. Evol.* **10**, 33–44 (2018).
24. G. J. Prideaux, *Systematics and Evolution of the Sthenurine Kangaroos. Univ. Calif. Publ. Geol. Sci. 146* (2004).
25. B. Pillans, R. Bourman, *Aust. J. Soil Res.* **39**, 89–98 (2001).
26. L. R. G. DeSantis, J. H. Field, S. Wroe, J. R. Dodson, *Paleobiology* **43**, 181–195 (2017).
27. G. J. Prideaux *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 11646–11650 (2009).
28. K. J. Mitchell *et al.*, *Mol. Biol. Evol.* **31**, 2322–2330 (2014).
29. C. H. Martin, P. C. Wainwright, *Science* **339**, 208–211 (2013).
30. J. Zachos, M. Pagani, L. Sloan, E. Thomas, K. Billups, *Science* **292**, 686–693 (2001).

Urbanization and humidity shape the intensity of influenza epidemics in U.S. cities

Benjamin D. Dalziel[1,2]*, Stephen Kissler[3], Julia R. Gog[3], Cecile Viboud[4], Ottar N. Bjørnstad[5], C. Jessica E. Metcalf[6,7], Bryan T. Grenfell[4,6,7]

Influenza epidemics vary in intensity from year to year, driven by climatic conditions and by viral antigenic evolution. However, important spatial variation remains unexplained. Here we show predictable differences in influenza incidence among cities, driven by population size and structure. Weekly incidence data from 603 cities in the United States reveal that epidemics in smaller cities are focused on shorter periods of the influenza season, whereas in larger cities, incidence is more diffuse. Base transmission potential estimated from city-level incidence data is positively correlated with population size and with spatiotemporal organization in population density, indicating a milder response to climate forcing in metropolises. This suggests that urban centers incubate critical chains of transmission outside of peak climatic conditions, altering the spatiotemporal geometry of herd immunity.

Predicting the epidemiology and evolution of influenza is an important goal for public health and an approaching milestone in the study of complex systems (1, 2). Patterns of influenza spread and diversification are shaped by interacting ecological and evolutionary processes, including viral antigenic evolution (3–6), climatic conditions affecting transmission potential (7), and spatial heterogeneity in transmission among hosts, from local (8, 9) to regional (10, 11) to global scales (12). A global latitudinal gradient in epidemic periodicity (with more strongly seasonal epidemics at temperate latitudes) is associated with climatic variation (13), with fluctuations in specific humidity as a key climatic driver (14). In temperate regions, multiyear "boom and bust" cycles in strain-specific incidence are associated with epochal evolution, involving intermittent jumps through antigenic space driven by antigenically localized susceptible depletion (3, 4). This process is coupled to an evolutionary backbone shaped by global migration patterns in the virus (15), including repeated seeding from persistent regions, particularly in Asia (16, 17). At regional scales, differences in epidemic timing are correlated with patterns of human contact, including commuting patterns and the timing of school terms (10, 11, 18, 19). Robust epidemic and antigenic forecasts require a predictive understanding of the emergent properties of these interacting processes (20, 21).

Cities are the principal locations for influenza transmission in humans (22), and therefore the primary context where drivers of transmission interact. However, recent comparisons of city-level influenza transmission patterns have revealed unexplained differences among cities within the same broad climatic and antigenic regimes, suggesting that endogenous differences among cities may interact with climatic and evolutionary drivers to cause divergent epidemic dynamics at the city level (23, 24). Cities can differ from each other in several ways that could potentially influence influenza transmission, including variation in the timing and coverage of public health interventions (25, 26) and variation in population health and socioeconomic conditions (27–29). Cities also differ fundamentally in population size, spatial structure, and connectivity, in ways that may affect infectious contact patterns (30, 31). These have the potential to substantially alter epidemic dynamics, including responses to climate forcing, and the impacts of public health interventions (28, 32, 33). However, the role of city size and structure in shaping transmission patterns of seasonal influenza is not well understood.

We address this here using 6 years (2002 to 2008) of data on weekly incidence of influenza-like illness (ILI) in 603 three-digit postal (ZIP) codes across the United States, assembled from medical claims data (34). ZIP codes are designed for efficient mail distribution such that the first three digits typically represent a contiguous geographic area surrounding a major city. Incidence in a ZIP code is measured as the proportion of physician visits that are for ILI in a given week, and is strongly correlated with U.S. Centers for Disease Control and Prevention (CDC) reference influenza surveillance time series (Spearman's

ρ > 0.88). Our analysis corrects for sensitivity and specificity of ILI surveillance as an estimate of influenza incidence by incorporating city-specific reporting rates that vary temporally between the peak and off-peak influenza season (materials and methods).

The ILI data show persistent differences among cities in how influenza incidence is distributed throughout each season (Fig. 1, A and B). Let the incidence distribution p_{ij} represent the fraction of ILI incidence in influenza year j (centered on Northern Hemisphere winter: 1 July to 30 June) that occurred during week i (weeks from 1 July of the current influenza year) in a given city, and define epidemic intensity, v_j, as the inverse of the Shannon entropy of incidence distribution in a given city and year, $v_j = \left(-\sum_i p_{ij} \log p_{ij}\right)^{-1}$, which we normalize to be between 0 and 1 by subtracting the global minimum and then dividing by the global maximum across all cities. Epidemic intensity thus defined is minimized when incidence is spread evenly across weeks and increases as incidence becomes more intensively focused in particular weeks. Because v_j is a function of incidence distribution, rather than raw incidence, it is invariant under differences in overall reporting rates and/or attack rates among cities and years.

We find that differences in v_j among cities persist across years—some cities have consistently more intense epidemics than others, year after year. These differences among cities are epidemiologically significant, comparable in magnitude to differences in intensity among years associated with antigenic shifts [e.g., the increase in epidemic intensity across all cities in the 2003–2004 season associated with the A/Fujian/411/02 (H3N2) strain variant is comparable in magnitude to differences in intensity among cities in any season; Fig. 1C], and are also apparent in separate publicly available data, and across wider time scales, including since the 2009 pandemic (29) (fig. S1).

Differences in mean intensity $v \propto \sum_j v_j$ show a geographic pattern, with intense epidemics focused in the east (Fig. 1D). Mean intensity also varies with population size (Fig. 1E) and climate (Fig. 1F). In particular, v tends to be higher in smaller populations, especially those with high amplitude in seasonal fluctuations of specific humidity.

We hypothesize that these patterns are caused by differential responses to climate forcing, mediated by divergent spatiotemporal patterns of transmission potential in cities of different sizes. By transmission potential, we mean the propensity for two randomly selected hosts in a population to attain spatiotemporal proximity sufficient for influenza transmission—sufficient proximity for the transfer of respiratory droplets from one host to the other. As specific humidity decreases in the winter, influenza virus remains viable outside a host for longer, expanding the spatiotemporal "cloud of risk" generated by an infected host and increasing transmission potential in the population. Seasonal, climate-driven

[1]Department of Integrative Biology, Oregon State University, Corvallis, OR, USA. [2]Department of Mathematics, Oregon State University, Corvallis, OR, USA. [3]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK. [4]Division of International Epidemiology and Population Studies, Fogarty International Center, National Institutes of Health, Bethesda, MD, USA. [5]Department of Entomology, Pennsylvania State University, State College, PA, USA. [6]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, USA. [7]Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton, NJ, USA.
*Corresponding author. Email: benjamin.dalziel@oregonstate.edu

increases in transmission potential thus drive the reproductive number of the infection (the expected number of secondary cases caused by an index case) upward in winter, eventually leading to an epidemic (*23*, *24*, *35*). However, climate is less important when the spatiotemporal distance between a pair of hosts is small, as is the case for a subset of potential contacts, such as those that reside, travel, or work in close proximity (fig. S2). This base transmission potential—transmission potential that is not strongly modulated by climate—could influence epidemic dynamics by facilitating influenza transmission over a wider range of climatic conditions, in turn reducing population-level susceptibility during the peak influenza season.

We thus propose that elevated base transmission potential in the presence of climate forcing leads to divergent epidemics among cities: Increased base transmission potential in urban centers enhances influenza spread outside of peak season, which elevates herd immunity to currently circulating strains, and subsequently attenuates explosive spread when climatic conditions are most favorable for transmission. This leads to the counterintuitive outcome that larger cities, with higher base transmission potentials, have more diffuse influenza epidemics. Base transmission potential may be elevated in large cities as a consequence of increased spatial organization, including aggregation of residences and work-

places, and the prevalence of high-density mass transit, among other factors (*30*, *31*).

We first demonstrate this effect using a standard climate-forced susceptible-exposed-infected-removed-susceptible (SEIRS) compartmental model for influenza epidemics (Fig. 2; materials and methods). Individuals enter the susceptible compartment in the model via immune waning following infection. New infections are generated by exposure of a susceptible individual to an infectious individual, at rate $\frac{\beta S I}{N}$, where N represents populations size, and $S(t)$ and $I(t)$ are functions of time that represent the number of susceptible and infectious individuals, respectively. For a given number of susceptible and infected individuals, the rate of appearance of new infections is governed by the transmission function $\beta(t) = \kappa + \sigma e^{-\omega q(t)}$, where κ represents city-level base transmission potential, σ the maximum gain in transmission potential at 0 specific humidity, and ω the rate of loss in viral viability due to specific humidity $q(t)$, in units of kg/kg. The transmission function $\beta(t)$ thus consists of a sum of two components: a seasonally invariant base transmission potential κ, representing transmission among contacts whose close spatiotemporal proximity renders climatic conditions moot; and additional transmission modulated by specific humidity, $\sigma e^{-\omega q(t)}$, which increases as drier conditions in U.S. cities in the winter increase the risk of transmission over larger spatiotemporal distances. This transmis-

sion function has been successfully used to fit and forecast seasonal influenza epidemics in previous studies (*14*, *24*).

The SEIRS model shown in Fig. 2 is proof of concept that increasing base transmission potential can decrease epidemic intensity, as predicted. However, there are several obstacles to confronting the model with incidence data in its current form. First, the model is a forced nonlinear oscillator, so small changes in parameter values may produce large changes in model predictions, which substantially complicates model fitting. Second, whereas the model assumes that incidence is perfectly observed in continuous time, the data consist of discrete (weekly) observations in each city, affected by city-specific time-varying differences in reporting rates. Finally, the model does not include interyear variation in transmission rates due to antigenic evolution.

Following previous work (*35–37*), we constructed a time-series approximation of the SEIRS model to work with city-level ILI data that accounts for variation in observed incidence driven by reporting and antigenic evolution (materials and methods). The resulting city-level time series models had 11 fitted parameters per city, i.e., <2 per year of incidence data, yet produced a strong match with the data, via *n*-week ahead (1 < *n* < 303) epidemic simulations (hindcasts; Fig. 3, A to C; Spearman's ρ = 0.93 for comparison of observed and model-predicted intensity). Out-of-sample simulation performance
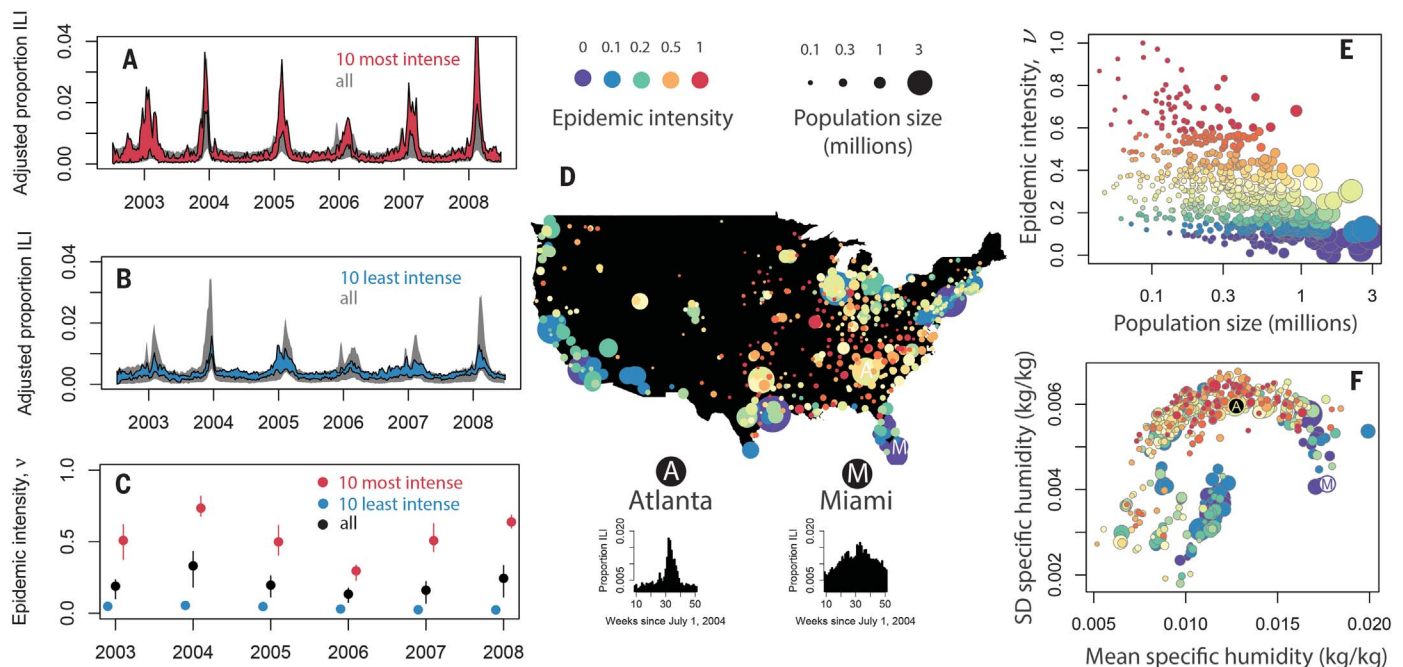
Fig. 1. Systematic differences among U.S. cities in the intensity of seasonal influenza epidemics. (**A** to **C**) Differences among cities in epidemic intensity are preserved across years, indicated by comparing the temporal dynamics of the cities with the highest and lowest average intensity. Points show means, vertical lines show interquartile ranges, and polygons enclose the central 95% of ILI incidence data that have been corrected for intercity variation in background incidence and reporting, by linear transformation of each city's time series to have minimum 0 and a common total attack rate over the 6-year period. (**D** to **F**) Cities with higher mean intensity tend to be located in the east, have smaller population sizes, and have higher-amplitude seasonal fluctuations in specific humidity. In (F), the vertical axis is standard deviation (SD) and the labeled points are Atlanta (A) and Miami (M).

was similarly strong (fig. S3). However, randomly reassigning κ estimates to cities destroyed the correspondence between the simulations and the data (fig. S4). Assessing the performance of *n*-week ahead time-series simulations requires comparing multiple features of the data and model predictions (*38*), and we also include time-series plots of observed and simulated data in each city in the supplementary materials.

Fitting the model to ILI time-series data from each city reveals that differences in κ, interacting with local patterns in specific humidity, are sufficient to explain observed differences in epidemic intensity among cities (Fig. 3, C, E, and F). Base transmission potential is correlated with overall population size in a city (N), with the average population size of a census block ($\bar{m}$), and with the level of crowding in each city (m^*) (Fig. 3, E and F, and fig. S5). Crowding is measured as the expected block-level population size experienced by a randomly selected individual within a city, $m^* = \bar{m} + \frac{\sigma_{\bar{m}}^2}{\bar{m}} - 1$ (*31*, *39*). As individuals within a city become aggregated within fewer focal locations, m^* increases above $\bar{m}$. We find that both m^* and $\bar{m}$ scale with city size (Fig. 3E), such that in large cities, residential and daytime population distribution are more highly organized. Moreover, circadian cycles of aggregation are more profound in large cities (steeper slope in daytime mean crowding compared to residential mean crowding as functions of population size in Fig. 3E: daytime slope = 0.537 ± 0.036, residential slope = 0.412 ± 0.023 SE).

Data on crowding substantially improve predictions of κ, relative to using population size alone, assessed using Akaike information criterion (AIC; ΔAIC = 37.36; Fig. 3F). Moreover, after adjusting for the effects of population size, residual crowding in cities is correlated with residual base transmission potential—cities that have more crowding for their size also have higher fitted values for κ than expected for their size ($p < 0.0001$ for linear regression of excess connectivity as a function of excess residential crowding and $p = 0.03$ for linear regression of excess connectivity as a function of excess daytime crowding)—consistent with the hypothesis that increased spatial organization in larger cities is driving increases in κ. Finally, interactions between κ and specific humidity provide a much stronger statistical fit to observed intercity variation in epidemic intensity, ν, relative to models featuring only specific humidity and/or population size (Fig. 3, D and G).

Regional correlations in seasonal influenza incidence have been linked with regional variation in city sizes and associated variation in intercity connectivity: All else being equal, random epidemic extinctions are less likely in large populations, and metropolises are more strongly interconnected by patterns of human travel, which synchronize epidemics among cities (*10*). However, influenza transmission dynamics within cities have generally been assumed to conform to
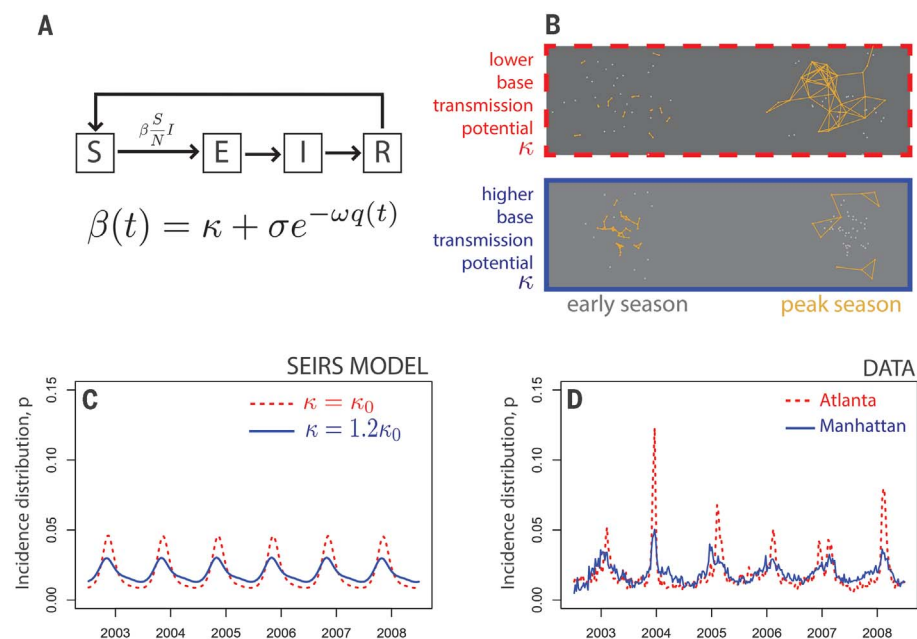
Fig. 2. Increasing base transmission potential can decrease epidemic intensity in a seasonally forced compartmental epidemic model. (**A**) Diagram of a susceptible-exposed-infected-removed-susceptible (SEIRS) model. The seasonally varying transmission rate β depends on specific humidity q and the base transmission potential of the population, κ. See materials and methods for details. (**B**) Diagram of transmission in two hypothetical populations. Points represent individual hosts and yellow lines show transmission events. In populations with higher base connectivity, chains of transmission are longer during the early influenza season, when climatic conditions are not yet ideal for wider spread. (**C**) Simulations of the model for two levels of base transmission (red and blue lines), which yield corresponding variation in epidemic intensity. (**D**) Incidence distributions in U.S. three-digit ZIP codes (e.g., Atlanta and Manhattan) show comparable variation in epidemic intensity, and also evidence of seasonal variation in transmission rates and reporting, which are included in the model during fitting (Fig. 3; materials and methods).

the assumptions of mass action, precluding systematic intercity differences in epidemic dynamics that are driven endogenously by differential contact patterns. By contrast, our results show that processes underlying epidemic persistence and interconnectivity are rooted at the intracity scale and drive divergent, yet highly predictable, responses to climate forcing among cities of different sizes, which then scale up to influence regional epidemic patterns. Because large cities are also hubs in the intercity travel network, spatial aggregation of populations in large cities could be a proxy for the intensity of infectious contact both within and among cities. In this context, a key uncertainty is how external seeding of infections may drive epidemic patterns among cities of different sizes, and more generally, how transmission processes within cities drive patterns in epidemic intensity at different scales of observation (*29*).

Our model predicts that changes in urbanization and climate will lead to specific changes in the intensity of future influenza epidemics. In particular, increasing the amplitude of seasonal fluctuations in specific humidity leads to more intense epidemics in our model; however, elevated base transmission potential in metropolises could

counteract this effect (fig. S6). Notably, vaccination early in the season could mimic the accumulation of population-level immunity via off-peak transmission, increasing both direct and indirect protection (*40*) and regulating the intensity of seasonal epidemics; this illustrates an additional population-level benefit to influenza vaccination under increasingly extreme climate cycles. At the same time, state-level variation in vaccination coverage is not associated with variation in epidemic intensity across cities (fig. S8), perhaps because yearly variation in vaccine efficacy dwarfs geographic differences in vaccine coverage, obscuring any residual effect of vaccination on the spatial patterning of epidemics (*29*, *40*).

The scale of influenza epidemics can sometimes mirror that of pandemics—for example, the recent influenza seasonal outbreak in winter 2017–2018 had a similar epidemic size and peak intensity as that of the 2009 pandemic in the United States. More research is needed to understand and predict the scale and intensity of influenza outbreaks, as a function of population susceptibility and spatial organization, and the potential trade-offs between these epidemic parameters. Our work indicates potential trade-offs between scale and intensity of epidemics that raise important
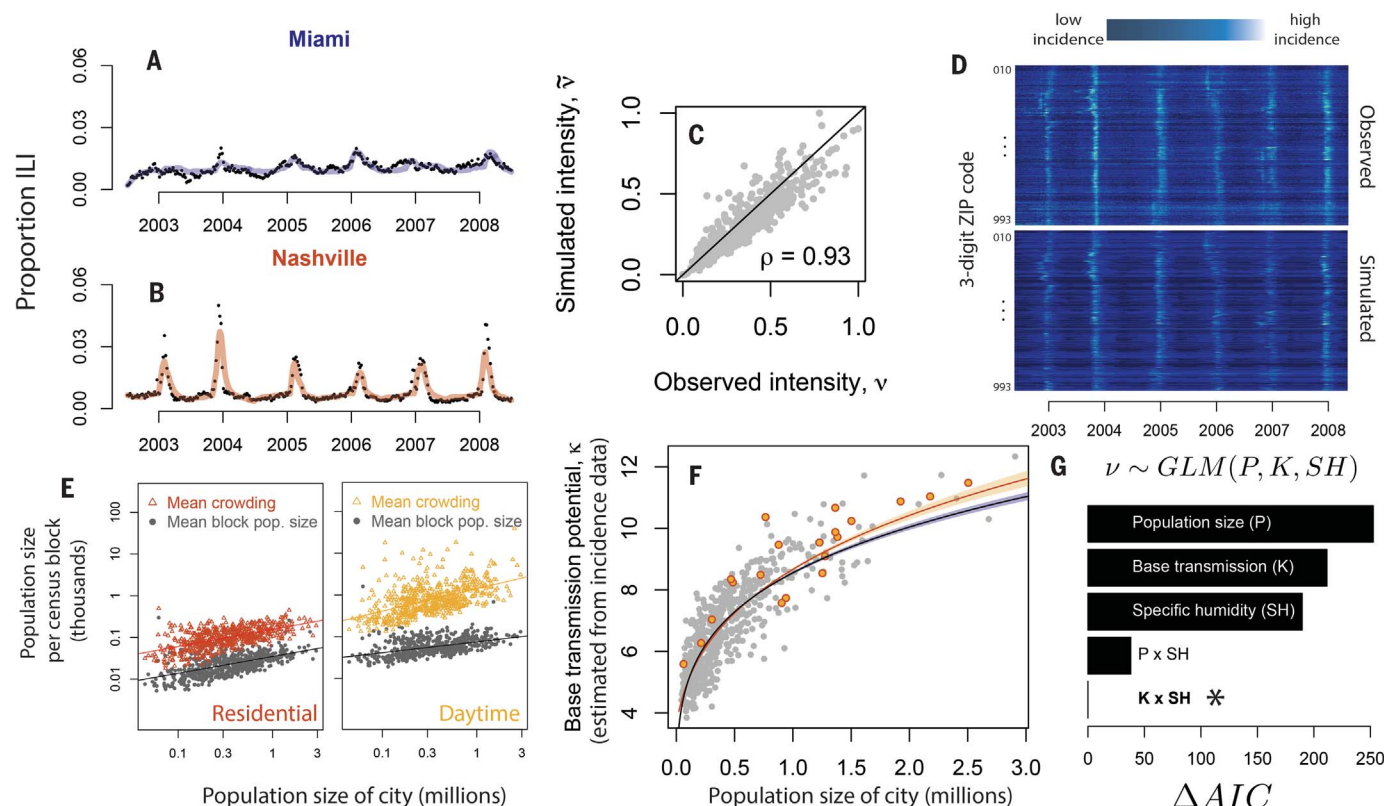
Fig. 3. Base transmission potential and specific humidity predict observed differences in the intensity of influenza epidemics across U.S. cities. (**A** and **B**) n-step ahead simulation performance of the fitted SEIRS model, $1 < n < 303$ weeks, in two cities. (**C**) Observed versus forward simulated average epidemic intensity in all cities. (**D**) Observed and simulated incidence in all cities and years. (**E**) Larger cities have more organized spatial population distributions and mobility patterns. Gray points show expected population size in a randomly selected census block in each city; colored points show expected block-level population size experienced by a randomly selected individual in each city [Lloyd's mean crowding $m^* = \bar{m} + \frac{\sigma_m^2}{\bar{m}} - 1$, where $\bar{m}$ represents mean population size in a census block, and σ_m^2 variance in population size across census blocks (31, 39)]. Mean crowding increases above mean block-level population size as spatial locations of individuals become more highly organized. (**F**) Population size and crowding estimated from census data predict base transmission potential estimated from ILI incidence data. Blue line shows fit for population size alone; yellow line, population size and crowding. Polygons enclose 1 SE around the fitted curves. Yellow points show the 20 cities with the most residential crowding. (**G**) Information-theoretic comparison of population size, climatic fluctuations, and fitted base transmission potential (i.e., base transmission potential predicted from population size and crowding rather than fitted to the incidence data) as predictors of observed epidemic intensity, via generalized linear models.

questions for future work on the optimization of health systems against endemic and pandemic threats.

Increased epidemic intensity demands increased surge capacity in the public health system, including primary care facilities and clinical laboratories (41). This is particularly important for influenza, where the impact of vaccination depends on timely development and distribution of annual vaccines (40). Our analysis shows that some of the cities with the most intense influenza epidemics (driven by low base transmission potentials) are also among those with the most challenging socioeconomic conditions (fig. S7) (27). This is congruent with recent analyses of socioeconomic determinants of influenza mortality at the intracity level (28). Statistical associations between socioeconomic conditions and influenza dynamics may thus be caused in part by underlying variation in human aggregation patterns: For instance, metropolises have

highly aggregated cores with high base transmission potentials, where median per-capita income is also higher. Our results also underscore the importance of considering spatial heterogeneity when assessing the impacts of climate forcing on infectious disease dynamics. As has recently been demonstrated for diarrheal diseases (33), spatial patterns in population density within cities can modulate the impact of climate variation on disease transmission patterns. By extending this result to include influenza, our findings indicate the potential for systematic effects of metropolises on climate-driven disease dynamics across a range of pathogens.

The ecological and evolutionary dynamics of influenza depend on the locations of "fertile ground" for transmission: places and times where critical chains of transmission incubate immigrating viral lineages (5, 42). Our results show how metropolises play a disproportionately important role in this process, as epidemic foci, and as

potential sentinel hubs, where epidemiological observatories could integrate local strain dynamics to predict larger-scale patterns (4, 43, 44). As the growth and form of cities affect their function as climate-driven incubators of infectious disease, it may be possible to design smarter cities that better control epidemics in the face of accelerating global change.

REFERENCES AND NOTES

1. S. Gandon, T. Day, C. J. E. Metcalf, B. T. Grenfell, *Trends Ecol. Evol.* **31**, 776–788 (2016).
2. D. H. Morris *et al.*, *Trends Microbiol.* **26**, 102–118 (2018).
3. K. Koelle, S. Cobey, B. Grenfell, M. Pascual, *Science* **314**, 1898–1903 (2006).
4. M. Łuksza, M. Lässig, *Nature* **507**, 57–61 (2014).
5. D. Zinder *et al.*, *BMC Evol. Biol.* **14**, 272 (2014).
6. T. Bedford *et al.*, *Nature* **523**, 217–220 (2015).
7. M. Lipsitch, C. Viboud, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 3645–3646 (2009).
8. S. Cauchemez *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 2825–2830 (2011).

9. L. Bourouiba, E. Dehandshoewercker, J. W. M. Bush, *J. Fluid Mech.* **745**, 537–563 (2014).
10. C. Viboud *et al.*, *Science* **312**, 447–451 (2006).
11. V. Charu *et al.*, *PLOS Comput. Biol.* **13**, e1005382 (2017).
12. M. Tizzoni *et al.*, *BMC Med.* **10**, 165 (2012).
13. J. D. Tamerius *et al.*, *PLOS Pathog.* **9**, e1003194 (2013).
14. J. Shaman, M. Kohn, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 3243–3248 (2009).
15. T. Bedford, S. Cobey, P. Beerli, M. Pascual, *PLOS Pathog.* **6**, e1000918 (2010).
16. C. A. Russell *et al.*, *Science* **320**, 340–346 (2008).
17. F. Wen, T. Bedford, S. Cobey, *Proc. Biol. Sci.* **283**, 20161312 (2016).
18. S. Cauchemez, A.-J. Valleron, P.-Y. Boëlle, A. Flahault, N. M. Ferguson, *Nature* **452**, 750–754 (2008).
19. J. R. Gog *et al.*, *PLOS Comput. Biol.* **10**, e1003635 (2014).
20. I. Chattopadhyay, E. Kiciman, J. W. Elliott, J. L. Shaman, A. Rzhetsky, *eLife* **7**, e30756 (2018).
21. S. Pei, S. Kandula, W. Yang, J. Shaman, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2752–2757 (2018).
22. United Nations, Department of Economic and Social Affairs, Population Division, "World Urbanization Prospects" (2014), pp. 1–32.
23. J. B. Axelsen, R. Yaari, B. T. Grenfell, L. Stone, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 9538–9542 (2014).
24. W. Yang, M. Lipsitch, J. Shaman, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 2723–2728 (2015).
25. R. J. Hatchett, C. E. Mecher, M. Lipsitch, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7582–7587 (2007).
26. E. M. Galarce, S. Minsky, K. Viswanath, *Vaccine* **29**, 5284–5289 (2011).
27. R. Chetty *et al.*, *JAMA* **315**, 1750–1766 (2016).
28. K. H. Grantz *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 13839–13844 (2016).
29. E. C. Lee *et al.*, *PLOS Comput. Biol.* **14**, e1006020 (2018).
30. L. M. A. Bettencourt, J. Lobo, D. Helbing, C. Kühnert, G. B. West, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7301–7306 (2007).
31. B. D. Dalziel, B. Pourbohloul, S. P. Ellner, *Proc. Biol. Sci.* **280**, 20130763 (2013).
32. B. D. Dalziel *et al.*, *PLOS Pathog.* **10**, e1004455 (2014).
33. P. P. Martinez, A. A. King, M. Yunus, A. S. G. Faruque, M. Pascual, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 4092–4097 (2016).
34. C. Viboud *et al.*, *PLOS ONE* **9**, e102429 (2014).
35. D. E. te Beest, M. van Boven, M. Hooiveld, C. van den Dool, J. Wallinga, *Am. J. Epidemiol.* **178**, 1469–1477 (2013).
36. O. N. Bjørnstad, B. F. Finkenstadt, B. T. Grenfell, *Ecol. Monogr.* **72**, 169–184 (2002).
37. B. D. Dalziel *et al.*, *PLOS Comput. Biol.* **12**, e1004655 (2016).
38. L. Held, S. Meyer, J. Bracher, *Stat. Med.* **36**, 3443–3460 (2017).
39. M. Lloyd, *J. Anim. Ecol.* **36**, 1–30 (1967).
40. N. Arinaminpathy *et al.*, *Am. J. Epidemiol.* **186**, 92–100 (2017).
41. J. M. Crawford *et al.*, *Emerg. Infect. Dis.* **16**, 8–13 (2010).
42. N. J. Hill *et al.*, *Emerg. Infect. Dis.* **23**, 654–657 (2017).
43. C. J. E. Metcalf *et al.*, *Lancet* **388**, 728–730 (2016).
44. K. S. Xue *et al.*, *eLife* **6**, e26875 (2017).

ACKNOWLEDGMENTS

Author contributions: Conceptualization: B.D.D.; Data curation: S.K., J.R.G., C.V.; Formal analysis: B.D.D.; Funding acquisition: C.J.E.M., B.T.G.; Investigation: B.D.D.; Methodology: B.D.D., S.K., J.R.G., O.N.B., C.V., C.J.E.M., B.T.G.; Software: B.D.D.; Visualization: B.D.D.; Writing, original draft: B.D.D.; Writing, review and editing: B.D.D., S.K., J.R.G., O.N.B., C.V., C.J.E.M., B.T.G.

SUPPLEMENTARY MATERIALS

FOREST ECOLOGY

Impacts of species richness on productivity in a large-scale subtropical forest experiment

Yuanyuan Huang[1]*, Yuxin Chen[1]*, Nadia Castro-Izaguirre[1], Martin Baruffol[1,2], Matteo Brezzi[1]†, Anne Lang[3], Ying Li[3], Werner Härdtle[3], Goddert von Oheimb[4,5], Xuefei Yang[6,7], Xiaojuan Liu[1,8], Kequan Pei[8], Sabine Both[6], Bo Yang[9], David Eichenberg[6,10], Thorsten Assmann[3], Jürgen Bauhus[11], Thorsten Behrens[12], François Buscot[4,13], Xiao-Yong Chen[14], Douglas Chesters[15], Bing-Yang Ding[16], Walter Durka[4,17], Alexandra Erfmeier[4,18], Jingyun Fang[19], Markus Fischer[20], Liang-Dong Guo[21], Dali Guo[22]†, Jessica L. M. Gutknecht[23], Jin-Sheng He[19], Chun-Ling He[15], Andy Hector[24], Lydia Hönig[6], Ren-Yong Hu[25], Alexandra-Maria Klein[26], Peter Kühn[12], Yu Liang[8], Shan Li[8], Stefan Michalski[17], Michael Scherer-Lorenzen[27], Karsten Schmidt[12], Thomas Scholten[12], Andreas Schuldt[3,4], Xuezheng Shi[28], Man-Zhi Tan[28], Zhiyao Tang[19], Stefan Trogisch[4,6,27], Zhengwen Wang[29], Erik Welk[4,6], Christian Wirth[4,10], Tesfaye Wubet[4,13], Wenhua Xiang[30], Mingjian Yu[31], Xiao-Dong Yu[15], Jiayong Zhang[32], Shouren Zhang[8], Naili Zhang[8], Hong-Zhang Zhou[15], Chao-Dong Zhu[15], Li Zhu[8], Helge Bruelheide[4,6]‡, Keping Ma[8]‡, Pascal A. Niklaus[1]‡, Bernhard Schmid[1]‡

Biodiversity experiments have shown that species loss reduces ecosystem functioning in grassland. To test whether this result can be extrapolated to forests, the main contributors to terrestrial primary productivity, requires large-scale experiments. We manipulated tree species richness by planting more than 150,000 trees in plots with 1 to 16 species. Simulating multiple extinction scenarios, we found that richness strongly increased stand-level productivity. After 8 years, 16-species mixtures had accumulated over twice the amount of carbon found in average monocultures and similar amounts as those of two commercial monocultures. Species richness effects were strongly associated with functional and phylogenetic diversity. A shrub addition treatment reduced tree productivity, but this reduction was smaller at high shrub species richness. Our results encourage multispecies afforestation strategies to restore biodiversity and mitigate climate change.

F orest ecosystems harbor around two-thirds of all terrestrial plant species. Observational studies suggest that species-rich forests exceed the productivity of less diverse forests (*1–3*), but covarying factors [such as spatial heterogeneity in abiotic environment (*1*), species composition (*2*), and successional stages (*2*)] make assigning causation difficult. Systematic experimental manipulations of plant species composition in grassland (*4–6*) have shown that plant diversity promotes community productivity through niche partitioning among species, specifically with respect to abiotic resources (*7*) or interactions with enemies (*8*), or through increasing the contribution of highly productive species in more diverse communities (*9*). These two types of biological mechanisms are thought to be captured by the complementarity and selection effects calculated by the additive partitioning of net biodiversity effects (*10*). Complementarity effects are large and positive when most species in a mixture contribute more than expected on the basis of their monoculture values to community values, and negative when most species in a mixture contribute less than expected, whereas selection effects are large when a single or few species show a disproportionate contribution to community values (*10*). It has been postulated that biodiversity effects may be weak or absent in forests, especially in those of high species richness, because the coexistence of so many species may require similar niches and competitive abilities (*1, 11–13*).

Several forest biodiversity experiments have recently been initiated (*14, 15*), mostly in the temperate zone or in small plots with limited species richness gradients (*16–22*). Here, we report results of the "BEF-China" experiment (BEF, biodiversity–ecosystem functioning) that was established in a highly diverse subtropical forest in southeast China (*23*). The experiment is characterized by a large species richness gradient, multiple simulated extinction scenarios, high replication, and extended duration (2009 to present). We studied how changing tree species richness affected stand-level development of tree basal area, aboveground volume, and aboveground carbon (C) from 2013 to 2017 (*24*). The experiment was implemented at two sites (site A and site B) of ~20 ha each, with communities assembled from six partially overlapping species pools (three per site). A complete pool represented a 16-species community, which was repeatedly divided to yield reduced richness levels of eight, four, two, and one species; in addition, 24-species communities were created by combining species of all three pools present at each site (fig. S1) (*24*). Of the 42 tree species used in the experiment (table S1), 40 occurred with the same frequency at each richness level. The remaining two species were typical plantation species in the area and were established in reference monocultures. A special feature of the design is that within each pool, communities form nested series that simulate different trajectories of trait-based species extinctions (fig. S2 and table S2). We analyzed trajectories related to

[1]Department of Evolutionary Biology and Environmental Studies, University of Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland. [2]Instituto de Investigación de Recursos Biológicos Alexander von Humboldt, calle 28A # 5-09, Bogotá DC, Colombia. [3]Institute of Ecology, Leuphana University of Lüneburg, Universitätsallee 1, 21335 Lüneburg, Germany. [4]German Centre for Integrative Biodiversity Research (iDiv) Halle–Jena–Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany. [5]Technische Universität Dresden, Institute of General Ecology and Environmental Protection, Post Office Box 1117, 01735 Tharandt, Germany. [6]Martin Luther University Halle-Wittenberg, Am Kirchtor 1, 06108 Halle (Saale), Germany. [7]Kunming Institute of Botany, Chinese Academy of Sciences, 134, Lanhei Road, Kunming, 650204, China. [8]State Key Laboratory of Vegetation and Environmental Change, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China. [9]Key Laboratory of Plant Resources and Biodiversity of Jiangxi Province, Jingdezhen University, 838 Cidu Avenue, Jingdezhen, Jiangxi 333000, China. [10]Institut für Spezielle Botanik und Funktionelle Biodiversität, University of Leipzig, 04103 Leipzig, Germany. [11]Chair of Silviculture, Faculty of Environment and Natural Resources, University of Freiburg, Tennenbacherstraße 4, 79106 Freiburg, Germany. [12]Department of Geosciences, Soil Science and Geomorphology, University of Tübingen, Rümelinstraße 19-23, 72074 Tübingen, Germany. [13]Helmholtz Centre for Environmental Research–UFZ, Department of Soil Ecology, Theodor-Lieser-Straße 4, 06120 Halle (Saale), Germany. [14]School of Ecological and Environmental Sciences, ECNU-UH Joint Translational Science and Technology Research Institute, East China Normal University, Shanghai 200241, China. [15]Institute of Zoology, Chinese Academy of Sciences, 1 Beichen West Road, Chaoyang District, Beijing 100101, China. [16]School of Life and Environment Sciences, Wenzhou University, Wenzhou, China. [17]Helmholtz Centre for Environmental Research–UFZ, Department of Community Ecology, Theodor-Lieser-Straße 4, 06120 Halle (Saale), Germany. [18]Institute for Ecosystem Research, Kiel University, Olshausenstraße 75, 24118 Kiel, Germany. [19]Department of Ecology, Peking University, 5 Yiheyuan Road, Beijing 100871, China. [20]Institute of Plant Sciences, University of Bern, Altenbergrain 21, 3013 Bern, Switzerland. [21]State Key Laboratory of Mycology, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China. [22]Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China. [23]Department of Soil, Water, and Climate University of Minnesota, Twin Cities, MN, USA. [24]Department of Plant Sciences, University of Oxford, South Parks Road, OX1 3RB, UK. [25]Department of Biology, College of Life and Environmental Sciences, Wenzhou University, Wenzhou 325035 China. [26]Nature Conservation and Landscape Ecology, Faculty of Earth and Environmental Sciences, University of Freiburg, Germany. [27]Geobotany, Faculty of Biology, University of Freiburg, Freiburg, Germany. [28]Institute of Soil Science, the Chinese Academy of Sciences, Nanjing 210008 China. [29]Institute of Applied Ecology, Chinese Academy of Sciences, 72 Wenhua Road, Shenyang 110016, P.R. China. [30]Faculty of Life Science and Technology, Central South University of Forestry and Technology, Changsha 410004, Hunan Province, China. [31]College of Life Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China. [32]Institute of Ecology, College of Chemistry and Life Science, Zhejiang Normal University, Yingbin Road No. 688, Jinhua City, Zhejiang Province, China 321004.
*These authors contributed equally to this work. †Deceased.
‡Corresponding author. Email: helge.bruelheide@botanik.uni-halle.de (H.B.); kpma@ibcas.ac.cn (K.M.); pascal.niklaus@ieu.uzh.ch (P.A.N.); bernhard.schmid@ieu.uzh.ch (B.S.)

means and diversities of the three functional traits leaf duration (LD), specific leaf area (SLA), and wood density (WD). These traits are often used to characterize plant-growth strategies (25) and are potentially related to extinction probabilities under environmental change (26). In 2009 (site A) and 2010 (site B), communities of 400 trees were planted on square plots 0.067 ha in size, which equals the Chinese area unit of 1 mu.

Communities of pools A2, A3, B2, and B3 (fig. S1) were established in single 1-mu plots. Each community of pools A1 and B1 was replicated in five 1-mu plots, four of which formed a larger square plot of 4 mu; these four plots received an understory shrub species richness treatment factorially crossed with the tree species richness gradient: Plots had zero, two, four, or eight shrub species randomly selected from a pool of 18

species, with shrubs planted at the same density as the trees.

We found significant positive effects of the logarithm of tree species richness on stand basal area and stand volume as well as on the annual increments of these two variables (Table 1, Fig. 1, and figs. S3 and S4). These effects grew steadily through time (changes in stand volume per doubling of species, with standard errors, were 0.74 ± 0.58, 1.47 ± 0.85, 2.98 ± 1.29, 4.91 ± 1.83, and 6.99 ± 2.24 m^3 ha^{-1} from 2013 to 2017). Mean volume increments were larger in wetter years ($F_{1,99.1} = 7.58$, $P = 0.007$), but richness effects on volume increments were not affected by annual precipitation ($F_{1,91.7} = 2.25$, $P = 0.137$). After 8 years of growth (site A), the average 16-species mixture stored 31.5 Mg C ha^{-1} above ground [95% Bayesian credible interval (CI), 25.5 to 37.6] (24), which is more than double the amount found in monocultures (11.9 Mg C ha^{-1}; CI, 10.6 to 13.5) (fig. S5) and similar to the C storage of monocultures of the commercial plantation species *Cunninghamia lanceolata* (26.3 Mg C ha^{-1}; CI, 19.0 to 33.2) and *Pinus massoniana* (28.5 Mg C ha^{-1}; CI, 20.8 to 36.1) (fig. S5). These strong positive effects of tree species richness must have been driven by faster growth of live trees in more diverse stands because tree survival rate did not increase with species richness (fig. S6). This is in contrast to findings in a large grassland biodiversity experiment in which positive diversity effects on productivity were mediated by a greater number rather than larger size of individuals in high-diversity plots (27).

The net biodiversity effect (10) on stand volume increased through time for mixtures of all species-richness levels (year as linear term with $F_{1,38.6} = 29.15$, $P < 0.001$) (Fig. 2) and was driven by increases in complementarity effects (year as linear term with $F_{1,52.4} = 9.23$, $P = 0.004$) (Fig. 2). Selection effects were on average negative ($F_{1,37.8} = 8.75$, $P = 0.005$) because some species with relatively high monoculture stand volume had lower performance in mixtures, and some with relatively low monoculture stand volume had higher performance. This was corroborated by negative species-level selection effects (fig. S7).

We tested whether the observed species-richness effects could be explained by functional or phylogenetic diversity. For this, we calculated functional diversity (FD) and functional dispersion (FDis) (24) on the basis of the seven plant
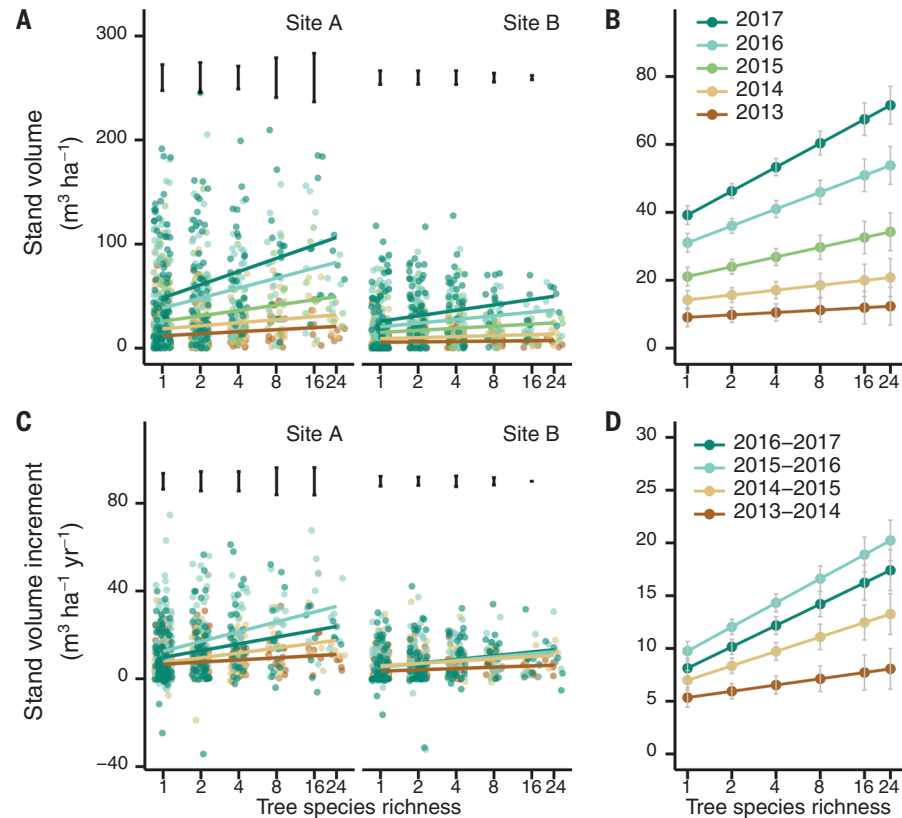
Fig. 1. Stand-level tree volume and its increment as a function of tree species richness from 2013 to 2017. (**A** and **B**) Stand-level tree volume. (**C** and **D**) Stand-level tree volume increment. In (A) and (C), raw data points and regression lines are shown for each year. (B) and (D) show predicted means and standard errors based on mixed-effects models (Table 1). The extremes of the point cloud taper off toward higher diversity levels because of decreasing sample size; quantile regressions show qualitatively the same positive relationships for the largest 10% of values at each diversity level (fig. S4). Standard deviations of species compositions (square root of corresponding between-composition variance components), shown as black error bars above the raw data, indicate that there is no variance-reduction effect of increasing species richness.

Table 1. Mixed-effects models for effects of site, tree species richness (logSR), year, and interactions on stand-level tree basal area and volume and their increments. Fixed effects were fitted sequentially (type-I sum of squares) as indicated in the table [random terms are provided in (24)]. *n*, numbers of plots; df, numerator degrees of freedom; ddf, denominator degrees of freedom; and logSR, log$_2$(tree species richness). F and *P* indicate F ratios and the *P* value of the significance test, respectively.

		Basal area (*n* = 387)				Volume (*n* = 387)				Basal area increment (*n* = 387)				Volume increment (*n* = 387)		
Source of variation	df	ddf	F	*P*	df	ddf	F	*P*	df	ddf	F	*P*	df	ddf	F	*P*
Site	1	120.0	14.35	<0.001	1	100.0	20.79	<0.001	1	121.5	8.12	0.005	1	101.3	20.79	<0.001
LogSR	1	111.9	7.45	0.007	1	88.9	6.62	0.012	1	113.8	15.58	<0.001	1	91.2	12.41	<0.001
Year	4	489.4	309.0	<0.001	4	402.3	197.10	<0.001	3	287.5	9.90	<0.001	3	281.8	35.05	<0.001
Site × year	4	488.3	7.75	<0.001	4	410.4	20.92	<0.001	3	301.0	9.43	<0.001	3	309.0	19.62	<0.001
LogSR × year	4	456.2	15.21	<0.001	4	368.9	11.98	<0.001	3	265.6	3.82	0.010	3	259.0	6.18	<0.001

Fig. 2. Changes over time in the net biodiversity effect (NE) and its additive components, complementarity effect (CE) and selection effect (SE), on stand-level tree volume in mixed-species plots. N = 65 to 77, 50 to 52, 28, and 14 plots for two-, four-, eight-, and 16-species mixtures, respectively. The figure shows means ± SEs. The y axis is square root–scaled to reflect the quadratic nature of biodiversity effects (10).
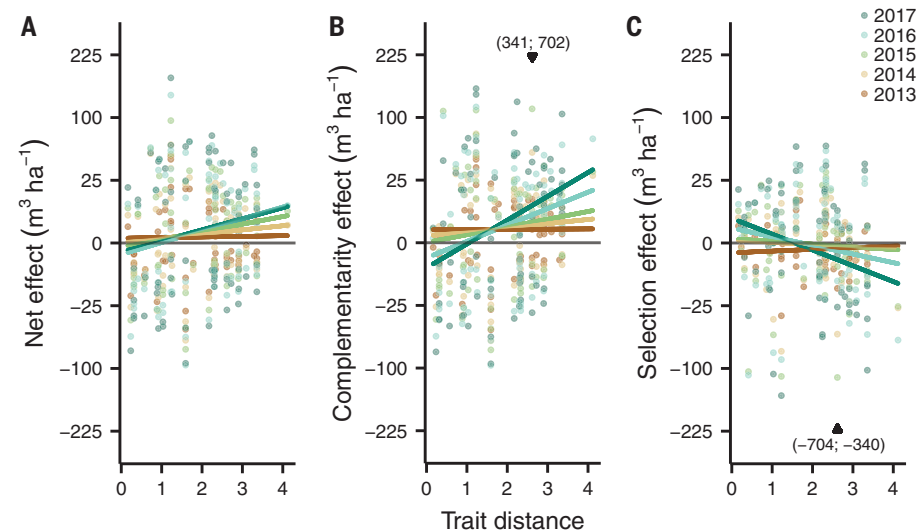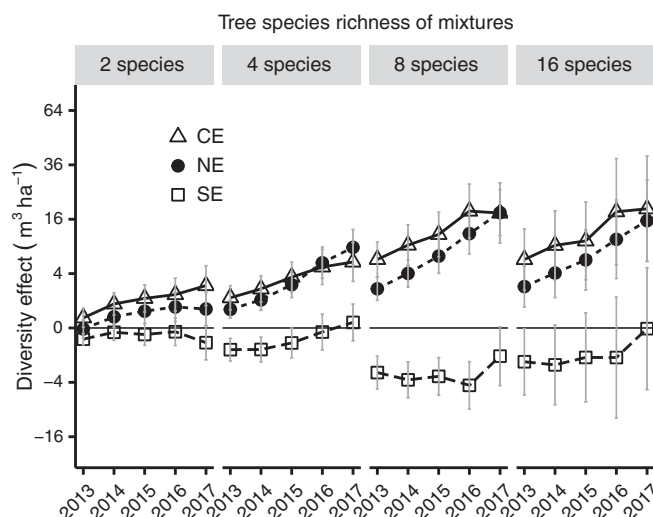
Fig. 3. Relationship between functional trait distance and biodiversity effects on stand volume in two-species mixtures across years. (**A** to **C**) Each point represents a plot in a year (n = 65 to 77 plots). Regression lines are based on mixed-effects models (24). Euclidean trait distances were calculated with the three z-transformed traits LD, SLA, and WD. The y axes are square root–scaled to reflect the quadratic nature of biodiversity effects (10). Two extreme y values are moved to the plot margin and given as numbers.

functional traits LD (deciduous or evergreen), SLA, WD, leaf dry matter content, leaf nitrogen, leaf phosphorus, and leaf thickness or the first three of these (LD, SLA, and WD), which contributed most to explanatory power. We also calculated phylogenetic diversity (PD) and mean phylogenetic distance (MPD) for each community (24). All measures of functional and phylogenetic diversity had similar explanatory power as that of species richness for stand-level productivity measures; differences between species-richness levels in stand volume could also be explained by associated differences in functional or phylogenetic diversity (fitted before species richness in model 1 in tables S3 and S4, respectively). However, none of the functional or phylogenetic diversity measures could explain additional variation among communities of the same richness level (when

fitted after species richness in model 2 in tables S3 and S4, respectively). This finding is consistent with similar reports from large-scale grassland biodiversity experiments (28). It is conceivable that for each particular species mixture with high stand-level productivity, a particular combination of functional traits causes the observed biodiversity effect; this cannot be captured by using the same functional diversity measure for all species mixtures.

Earlier studies have suggested that positive biodiversity effects in forests might originate from denser crown packing and enhanced light interception in mixed-species canopies (21, 29, 30). We measured the vertical crown extent of all trees in 2016 and 2017 and tested whether plots with less crown overlap produced greater stand-level volume (24), which was not the case ($F_{1,446.8}$ = 1.73,

P = 0.189). A reason for the absence of such a correlation might be that depending on the particular species combination, crown dissimilarity can result from light competition (18) or from complementary light use among species.

Despite the absence of general effects of functional diversity beyond species richness, we found some specific effects along the multiple extinction scenarios inherent in our experimental design (fig. S2A) (24). Changes in FD with each halving of species richness were negatively correlated with stand-volume changes at high but positively correlated at low species richness (fig. S8A), suggesting that FD captured beneficial differences between species at low but not at high diversity. We then focused on mixtures of two species because for these, the highest number of distinct species compositions were available. We found that a positive correlation of net biodiversity and complementarity effects with functional-trait distances developed over the 5 years of measurements (Fig. 3 and table S5). This was also the case for the diversity of the trait LD, indicating that mixtures of a shade-tolerant evergreen and a shade-intolerant deciduous broad-leaved species captured more light than did species pairs with uniform leaf duration.

Extinction sequences that differed in trajectories of community-weighted means for LD, SLA, or WD (fig. S2, B to D) did not show any significant variation in species-richness effects on stand-level productivity (fig. S8, B to D). This suggests that effects of trait-based extinctions, at least the ones tested and often considered most important (25, 26), may not differ much from effects of random extinction. Different results might have been obtained with other trait-based extinction scenarios, either ones that we did not analyze (for example, based on root traits) or ones that we did not simulate.

Plots additionally planted with shrubs (24) had reduced stand-level tree volume ($F_{1,234.5}$ = 7.30, P = 0.007), which is consistent with other findings that shrub removal in forests can increase tree growth (31). However, the effect of shrub competition decreased with increasing shrub species richness (log shrub richness $F_{1,191.9}$ = 6.57, P = 0.011), even though stand-level basal area of shrubs did not decrease (fig. S9). The reduced competition between shrubs and trees at higher shrub diversity suggests that complementarity effects extend to tree interactions with shrubs.

Our results provide strong evidence for a positive effect of tree species richness on tree productivity at stand level in establishing subtropical forest ecosystems and support the idea that co-occurring species in highly diverse subtropical forest can differ in niches and competitive abilities. At the end of the observation period, mixed stands with 16 species had accumulated about 1.7 times the amount of C found in the average monoculture (fig. S5). This effect is, on a relative scale, similar to the 1.8-fold average increase in aboveground stand biomass from monocultures to 16-species mixtures in a multisite grassland biodiversity experiment (4). Given that plant biomass is higher in forests, and that the largest

fraction of tree C is bound in relatively persistent woody biomass, these effects translate into large diversity-mediated rates of C accumulation. Specifically, after 8 years of growth at site A, we found an extra 19.5 (95% CI, 14.1 to 25.1) Mg C ha^{-1} accumulated in 16-species mixtures relative to the average monoculture. The biodiversity-productivity effects that we found did not differ between 1-mu and 4-mu plots ($F_{1,118.5} = 0.07$, $P > 0.5$ for interaction log tree species richness × plot size). However, biodiversity effects might be even larger at spatial scales beyond the ones that we tested experimentally because environmental heterogeneity might promote spatial insurance effects (32). Our first-order extrapolation to the global scale indicated that a 10% decrease in tree species richness would lead to a 2.7% decrease in forest productivity on average (24), which is within the range of productivity decreases (2.1 to 3.1%) reported for the same tree species loss scenario in a recent observational study that used plot data covering a large part of the global forests (3). In that study, it was estimated that such a loss would correspond to around $20 billion per year of commercial wood production.

Substantial forest areas are managed worldwide, with large afforestation programs underway (33, 34); in China, the total forested area increased by 1.5×10^6 ha year^{-1} from 2010 to 2015, mainly because of new monoculture plantation of species with high short-term productivity (35). Our experimental findings suggest that a similar or potentially even higher productivity can be achieved with mixed plantations of native species. Such strategies would yield cobenefits (2) in terms of active biodiversity management and likely higher levels of stability of productivity and ecosystem services under adverse conditions such as pathogen infestation or future climate change, including extreme events.

REFERENCES AND NOTES

1. A. S. Mori, *J. Ecol.* **106**, 113–125 (2018).
2. L. Gamfeldt *et al.*, *Nat. Commun.* **4**, 1340 (2013).
3. J. Liang *et al.*, *Science* **354**, eaaf8957 (2016).
4. A. Hector *et al.*, *Science* **286**, 1123–1127 (1999).
5. P. B. Reich *et al.*, *Science* **336**, 589–592 (2012).
6. D. Tilman, D. Wedin, J. Knops, *Nature* **379**, 718–720 (1996).
7. D. Tilman, C. L. Lehman, K. T. Thomson, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 1857–1861 (1997).
8. S. A. Schnitzer *et al.*, *Ecology* **92**, 296–303 (2011).
9. M. A. Huston, *Oecologia* **110**, 449–460 (1997).
10. M. Loreau, A. Hector, *Nature* **412**, 72–76 (2001).
11. H. Bruelheide *et al.*, *Ecol. Monogr.* **81**, 25–41 (2011).
12. S. P. Hubbell, *Ecology* **87**, 1387–1398 (2006).
13. X. Wang *et al.*, *Ecology* **97**, 347–360 (2016).
14. M. Scherer-Lorenzen, "The functional role of biodiversity in the context of global change" in *Forests and Global Change*, D. A. Coomes, D. F. R. P. Burslem, W. D. Simonson, Eds. (Cambridge Univ. Press, 2014), pp. 195–238.
15. K. Verheyen *et al.*, *Ambio* **45**, 29–41 (2016).
16. J. J. Grossman, J. Cavender-Bares, S. E. Hobbie, P. B. Reich, R. A. Montgomery, *Ecology* **98**, 2601–2614 (2017).
17. C. Potvin, N. J. Gotelli, *Ecol. Lett.* **11**, 217–223 (2008).
18. J. Sapijanskas, A. Paquette, C. Potvin, N. Kunert, M. Loreau, *Ecology* **95**, 2479–2492 (2014).
19. C. M. Tobner *et al.*, *Ecol. Lett.* **19**, 638–647 (2016).
20. T. Van de Peer, K. Verheyen, Q. Ponette, N. N. Setiawan, B. Muys, *J. Ecol.* **106**, 1096–1105 (2017).
21. L. J. Williams, A. Paquette, J. Cavender-Bares, C. Messier, P. B. Reich, *Nat. Ecol. Evol.* **1**, 63 (2017).
22. D. A. Clarke, P. H. York, M. A. Rasheed, T. D. Northfield, *Trends Ecol. Evol.* **32**, 320–323 (2017).
23. H. Bruelheide *et al.*, *Methods Ecol. Evol.* **5**, 74–89 (2014).
24. Materials and methods are available as supplementary materials.
25. P. B. Adler, A. Fajardo, A. R. Kleinhesselink, N. J. B. Kraft, *Ecol. Lett.* **16**, 1294–1306 (2013).
26. S. Greenwood *et al.*, *Ecol. Lett.* **20**, 539–553 (2017).
27. E. Marquard *et al.*, *J. Ecol.* **97**, 696–704 (2009).
28. D. F. B. Flynn, N. Mirotchnick, M. Jain, M. I. Palmer, S. Naeem, *Ecology* **92**, 1573–1581 (2011).
29. T. Jucker, O. Bouriaud, D. A. Coomes, *Funct. Ecol.* **29**, 1078–1086 (2015).
30. P. A. Niklaus, M. Baruffol, J.-S. He, K. Ma, B. Schmid, *Ecology* **98**, 1104–1116 (2017).
31. R. G. Wagner, K. M. Little, B. Richardson, K. McNabb, *Forestry* **79**, 57–79 (2006).
32. F. Isbell *et al.*, *Nature* **546**, 65–72 (2017).
33. Food and Agriculture Organization of the United Nations (FAO), *Global Forest Resources Assessment* (FAO, 2015).
34. R. J. Keenan *et al.*, *For. Ecol. Manage.* **352**, 9–20 (2015).
35. F. Hua *et al.*, *Nat. Commun.* **7**, 12717 (2016).

ECONOMICS

The role of education interventions in improving economic rationality

Hyuncheol Bryant Kim[1]*, Syngjoo Choi[2]*, Booyuel Kim[3]*, Cristian Pop-Eleches[4]*

Schooling rewards people with labor market returns and nonpecuniary benefits in other realms of life. However, there is no experimental evidence showing that education interventions improve individual economic rationality. We examine this hypothesis by studying a randomized 1-year financial support program for education in Malawi that reduced absence and dropout rates and increased scores on a qualification exam of female secondary school students. We measure economic rationality 4 years after the intervention by using lab-in-the-field experiments to create scores of consistency with utility maximization that are derived from revealed preference theory. We find that students assigned to the intervention had higher scores of rationality. The results remain robust after controlling for changes in cognitive and noncognitive skills. Our results suggest that education enhances the quality of economic decision-making.

Rationality in human choices has been a cornerstone assumption in traditional economic analysis and yet one of the most controversial issues in social and behavioral sciences (*1*). Mounting evidence shows that people tend to make systematic errors in judgment and decision-making and that there is a high level of heterogeneity in the extent to which rationality is limited across decisions and individuals (*2, 3*). The welfare loss resulting from poor decisions can be substantial, which implies that policy-makers might want to rethink the role of public policy in response to the failure of rationality (*4*).

The behavioral science literature has accumulated evidence on ways of improving people's capabilities and quality of decision-making: changing incentives, restructuring choice architecture, and debiasing training (*5–7*). Most of these approaches target the reduction of decision biases in particular contexts of economic activities but do not address the improvement of general capabilities of decision-making that are transferrable across decision domains. It is often controversial to judge whether decision biases are driven by the failure of rationality or other factors such as anomalous preferences.

Schooling has been shown to influence a wide range of outcomes, including income, health, and crime (*8, 9*). One little-explored hypothesis is that education improves people's decision-making abilities and leads them to make better decisions across various choice environments. The impacts of education on decision-making can then be a potential mechanism underlying the pecuniary and nonpecuniary returns to education.

We examine this hypothesis by studying a nongovernmental organization–implemented randomized controlled trial of education support in Malawi, an environment where, among young females, only 21.4% have received some secondary education and 9.8% have completed secondary school education (*10*). The program randomly provided financial support for education in a sample of 2812 female 9th and 10th graders from 83 classrooms in 33 public schools between the third semester of the academic year 2011–2012 and the second semester of the academic year 2012–2013. The program was randomized at the classroom level and consisted of the payment of school tuition and fees for 1 year, as well as a monthly cash stipend. The total amount of support was ~$70 per student as long as the student remained in school until the end of the program.

We conducted a short-term follow-up survey about 1 year later that measured short-term educational impacts. Four years after the intervention, we conducted a long-term follow-up survey that measured longer-term educational outcomes and implemented laboratory experiments of presenting subjects with a set of decision problems under risk and over time using a two-dimensional budget set. The risk-domain experiment consists of 20 decision problems representing a set of portfolio options associated with two equally probable unknown states. The time-domain experiment consists of two frames wherein the budget set represents a set of money allocations between two payment dates. The near time frame comprises 15 decisions of allocating money between tomorrow and 31 days from the time of the experiment. The distant time frame consists of 15 decisions of allocating money between 1 year and 1 year and 30 days from the time of the experiment.

This tool of laboratory experiments generates a rich set of individual choice data that are well suited to testing for consistency with utility maximization as the criterion for economic rationality (*3, 11, 12*). Classical revealed preference theory shows that choices from a finite collection of budget lines are consistent with maximizing a (well-behaved) utility function if and only if they satisfy the Generalized Axiom of Revealed Preference (GARP) (*13*). When the choice data do not satisfy GARP, we use Afriat's critical cost efficiency index (CCEI) to measure how closely they comply with the utility maximization hypothesis (*14*). We compute CCEI in each experimental domain to generate an index of the subject's level of economic rationality. The CCEI is bounded between 0 and 1. The closer the CCEI is to 1, the more closely the choice data are consistent with utility maximization. As the summary of economic rationality indices for the time domain, we use the minimum of two CCEIs at the near and distant time frames. The reason why consistency with utility maximization may be key to economic survival and thus serve as the basic criterion of economic rationality is offered by the classic money pump argument, which shows that inconsistent behavior can be exploited indefinitely by arbitrageurs. In addition, we consider two measures of compliance with stochastic dominance in the risk-domain experiment as an alternative criterion for economic rationality. Further details of the education intervention, laboratory experiment, and measurements are reported in the supplementary materials (*15*).

We present coefficients from regressions with baseline controls consisting of individual characteristics, parents' education and occupation, and school type. We cluster our standard errors at the classroom level. Because we deal with multiple outcomes of education and economic rationality, as well as the heterogeneous effects for 9th and 10th graders, we account for multiple hypothesis testing by following the approach in our preanalysis plan (*16*). We group all outcomes for the whole sample, 9th graders, and 10th graders in each realm of education or economic rationality and report standardized treatment effects with baseline controls as in (*17*), as well as family-wise adjusted P values.

First we evaluate the impacts of the intervention on various education outcomes: number of days absent during the past semester, school dropout rate, taking the Junior Certificate Examination (JCE) in 10th grade, passing the JCE, and total years of education. The information on the JCE comes from administrative data, whereas absence, dropout rate, and years of education are self-reported. Results for the whole sample, 9th graders, and 10th graders are presented in Table 1.

Students in the treated classrooms (i.e., those that were assigned the intervention) have better education outcomes compared with those in the control classrooms. Specifically, the treated students in the whole sample are 40% (1.6 days) less likely to be absent and 7% (5.5 percentage points) and 14% (8.6 percentage points) more likely to take and pass the JCE, respectively. The self-reported dropout rate decreases by 3.4 percentage points

[1]Department of Policy Analysis and Management, Cornell University, Ithaca, NY, USA. [2]Department of Economics, Seoul National University, Seoul, Republic of Korea. [3]KDI School of Public Policy and Management, Sejong, Republic of Korea. [4]School of International and Public Affairs, Columbia University, New York, NY, USA.
*Corresponding authors. Email: hk788@cornell.edu (H.B.K.); syngjooc@snu.ac.kr (S.C.); bkim@kdischool.ac.kr (B.K.); cp2124@columbia.edu (C.P.-E.)

Table 1. Impacts of education support program on education outcomes. Coefficients are from linear regressions of each education outcome on the education intervention indicator. Standard errors (in parentheses) are clustered at the classroom level. FU, follow-up; N/A, not applicable.

Sources:	Short-term FU survey	Short-term FU survey	Administrative data	Administrative data	Long-term FU survey	Combined data
Variables:	Absence, self-reported (2013)	Dropout, self-reported (2013)	Took JCE (2012–2013)	Passed JCE (2012–2013)	Total years of education (2015–2016)	Standardized treatment effect
	1	2	3	4	5	6
Overall sample						
Treated	−1.612***	−0.034	0.055**	0.086**	0.103	0.026***
	(0.397)	(0.026)	(0.023)	(0.033)	(0.076)	(0.008)
Family-wise adjusted *P* values	0.001	0.231	0.052	0.036	0.190	N/A
Control group mean	4.01	0.112	0.789	0.597	11.5	0
Number of observations	1851	1929	2808	2808	2420	11,816
Baseline 9th graders						
Treated	−1.498***	−0.083**	0.123***	0.141***	0.147	0.042***
	(0.343)	(0.039)	(0.040)	(0.049)	(0.119)	(0.013)
Family-wise adjusted *P* values	0.001	0.052	0.013	0.021	0.231	N/A
Control group mean	3.53	0.135	0.624	0.509	11.3	0
Number of observations	855	889	1220	1220	1051	5235
Baseline 10th graders						
Treated	−1.761**	−0.004	−0.004	0.043	0.058	0.012
	(0.703)	(0.033)	(0.021)	(0.040)	(0.098)	(0.009)
Family-wise adjusted *P* values	0.052	0.945	0.945	0.320	0.740	N/A
Control group mean	4.28	0.099	0.888	0.649	11.6	0
Number of observations	996	1040	1588	1588	1369	6581

**P < 0.05;   ***P < 0.01.

(30%), and the total years of education increases by ~0.1 of a year; however, neither change is statistically significant, and we expect a lot of measurement error in these self-reported measures. These treatment effects are heterogeneous between the two cohorts and come mainly from 9th graders: Those treated 9th graders are 42% (1.5 days) less likely to be absent per semester, 61% (8.3 percentage points) less likely to have dropped out, and 20% (12.3 percentage points) and 28% (14.1 percentage points) more likely to take and pass the JCE, respectively.

The standardized treatment effect shown in column 6 of Table 1 confirms that the intervention was successful in enhancing schooling, and this result is mainly driven by 9th graders. We interpret this heterogeneity because 9th graders are more vulnerable to dropping out of school and therefore could benefit more from the education intervention than 10th graders. Using data from our study, we estimate that the dropout rate was 26.4% in 9th grade compared with only 11.2% in 10th grade. This pattern is consistent with data from Malawian national statistics, as well as other settings (*15*).

Next we study whether the educational intervention affected economic rationality. Columns 1 and 2 of Table 2 present the average treatment effects on the CCEIs in the risk and time domains with baseline controls. For the overall sample, we observe that girls who received the intervention display an increase in

CCEIs measured in the risk and time domain of 1.3 percentage points (1.6%) and 1.4 percentage points (1.7%), respectively, but only the time domain effect is significant at the 5% level. Turning to the alternative measure of economic rationality, both the relative frequency and the expected payoff ratio of complying with stochastic dominance exhibit similar patterns as shown in columns 3 and 4. The standardized treatment effect across all four measures indicates that the treatment is associated with a 0.02 standard deviation [standard error (SE) = 0.009] increase in economic rationality scores. Figure S2 shows that the intervention improves economic rationality throughout most of the range of CCEIs.

Table 2 confirms that the intervention had heterogeneous impacts. For 9th graders in the control group, the mean CCEIs measured by risk and time domain are 0.81 and 0.82, respectively. The CCEIs of the treatment group are 3.3 percentage points (4.0%) and 3.1 percentage points (3.7%) higher than those of the control group (columns 1 and 2, respectively). Those treated among 9th graders are more likely to make decisions in conformity with stochastic dominance than those in the control group (columns 3 and 4). The standardized treatment effect for 9th graders across all measures is 0.038 standard deviations (SE = 0.011) (column 5). We do not find any treatment effect for 10th graders, but the statistical significance of the treatment

effects on economic rationality remain robust when using family-wise adjusted *P* values to account for multiple hypothesis testing.

A natural question that arises is whether our measures of economic rationality are proxies that are correlated with other primitives of decision-making that might also be affected by the intervention. To address this issue, we first examined the treatment effects of the intervention on time and risk preferences, cognitive abilities, and personality (*15*) and found that the intervention did not affect risk attitudes and time impatience but did enhance cognitive skills measured by the math test score and some aspects of personality traits (table S5). We then investigated the effects of the intervention on economic rationality, controlling for measures of risk and time preferences, cognitive skills, and personality. Our rationality scores are explained only partially by these control variables (table S6): For example, for 9th graders, the control variables reduce the impacts on rationality scores by about one-third.

There could be other explanations for our findings. First, the intervention might help subjects better understand the experiment instructions. However, when we drop the first three choices in the experiments, our results are robust, which suggests that differential learning during the experiment is not important (table S7). Second, beneficiaries might exert differential effort during the experimental games, despite

Table 2. Impacts of education support program on economic rationality.
Coefficients are from linear regressions of each rationality measure on the education intervention indicator. Standard errors (in parentheses) are clustered at the classroom level. N/A, not applicable.

Variables	CCEI, risk domain	CCEI, time domain	Compliance with stochastic dominance		Standardized treatment effect
			Freq.	Payoff	
	1	2	3	4	5
Overall sample					
Treated	0.013	0.014**	0.012*	0.005*	0.020**
	(0.008)	(0.006)	(0.007)	(0.003)	(0.009)
Family-wise adjusted *P* value	0.084	0.018	0.070	0.080	N/A
Control group mean	0.81	0.82	0.83	0.94	0.00
Number of observations	2421	2416	2421	2421	9679
Baseline 9th graders					
Treated	0.033***	0.031***	0.018**	0.009**	0.038***
	(0.010)	(0.008)	(0.008)	(0.004)	(0.011)
Family-wise adjusted *P* value	<0.001	<0.001	0.025	0.025	N/A
Control group mean	0.83	0.83	0.84	0.94	0.00
Number of observations	1051	1050	1051	1051	4203
Baseline 10th graders					
Treated	−0.003	0.003	0.006	0.002	0.005
	(0.010)	(0.009)	(0.010)	(0.004)	(0.012)
Family-wise adjusted *P* value	0.871	0.871	0.765	0.862	N/A
Control group mean	0.80	0.82	0.82	0.93	0.00
Number of observations	1370	1366	1370	1370	5476

*$P < 0.10$; **$P < 0.05$; ***$P < 0.01$.

the fact that they are incentivized. We created several measures aimed at capturing effort during the survey, including indexes of missing and "do not know" responses and did not find differences between the treatment and control groups (table S8). Third, we cannot positively distinguish whether the intervention improved rationality only via its effect through increased education. For example, the monthly stipend that was part of the intervention could lead girls to think more rationally about how to spend their money.

Using a randomized controlled trial of education support and financially incentivized laboratory experiments, we established causal evidence that an education intervention increases not only educational outcomes but also economic rationality. The size of the treatment effects on CCEIs is economically meaningful and larger than the cross-sectional relationship between education and CCEIs in our control group and the study from the Netherlands (*3*), as well as recent work on this relationship using changes in compulsory schooling in England (*18*). The direct comparison of the results between our study and the two aforementioned studies (*3, 18*) is difficult for several reasons. First, measures of years of schooling coming from self-reported levels of education achievement are generally noisy (*8*), especially so in a developing country setting where drop-out and grade repetition are frequent (*19*). Second, there are differences in laboratory experimental design such as the num-

ber of choices per subject and the variations of budget sets. Third, our treatment effects are measured after 4 years, whereas the other studies (*3, 18*) measure outcomes during adulthood. If program effects fade out over time, they could help reconcile the different results in these three studies (*20*). Fourth, the English and Dutch samples differ from our sample along many dimensions of socioeconomic status, and therefore our findings might not apply to populations in developed countries. For example, people in developed countries may have more opportunities to learn to make more rational decisions outside of school. Finally, on a hopeful note, our relatively larger impacts of education interventions are consistent with the literature that shows larger returns to cognitive and noncognitive investments in resource constraint settings (*21, 22*). In our setting, the impact of the educational intervention is large, not just in terms of the effects on economic rationality but also on cognitive outcomes (*15*). More research is needed to determine the reproducibility and generalizability of our findings.

REFERENCES AND NOTES

1. D. Kahneman, *Am. Econ. Rev.* **93**, 1449–1475 (2003).
2. D. Kahneman, P. Slovic, A. Tversky, *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge Univ. Press, 1982).
3. S. Choi, S. Kariv, W. Müller, D. Silverman, *Am. Econ. Rev.* **104**, 1518–1550 (2014).
4. R. H. Thaler, C. R. Sustein, *Nudge: Improving Decisions About Health, Wealth, and Happiness* (Penguin Books, 2008).
5. K. G. Volpp *et al.*, *N. Engl. J. Med.* **360**, 699–709 (2009).
6. R. H. Thaler, S. Benartzi, *J. Polit. Econ.* **112** (suppl. 1), S164–S187 (2004).
7. C. K. Morewedge *et al.*, *Policy Insights Behav. Brain Sci.* **2**, 129–140 (2015).
8. D. Card, in *Handbook of Labor Economics*, vol. 3A (Elsevier Science, 1999), chap. 30, pp. 1801–1863.
9. P. Oreopoulos, K. G. Salvanes, *J. Econ. Perspect.* **25**, 159–184 (2011).
10. National Statistical Office (NSO) Malawi, ICF, "Malawi Demographic and Health Survey 2015-16" (NSO and ICF, 2017); https://dhsprogram.com/publications/publication-fr319-dhs-final-reports.cfm.
11. S. Choi, R. Fisman, D. Gale, S. Kariv, *Am. Econ. Rev.* **97**, 1921–1938 (2007).
12. R. Fisman, P. Jakiela, S. Kariv, D. Markovits, *Science* **349**, aab0096 (2015).
13. S. N. Afriat, *Int. Econ. Rev.* **8**, 67–77 (1967).
14. S. N. Afriat, *Int. Econ. Rev.* **13**, 568–598 (1972).
15. See supplementary materials.
16. S. Choi *et al.*, "The Impacts of Female Education: Evidence from Malawian Secondary Schools," AEA RCT Registry ID AEARCTR-0001243 (2016); www.socialscienceregistry.org/trials/1243J.
17. R. Kling, J. B. Liebman, L. F. Katz, *Econometrica* **75**, 83–119 (2007).
18. J. Banks, L. S. Carvalho, F. Perez-Arce, "Education, decision-making, and economic rationality" (CESR-Schaeffer Working Paper Series no. 2018-003, 2018); https://cesr.usc.edu/documents/WP_2018_003.pdf.
19. P. Glewwe, M. Kremer, in *Handbook of the Economics of Education* (Elsevier, 2006), vol. 2, pp. 945–1017.
20. J. Heckman, R. Pinto, P. Savelyev, *Am. Econ. Rev.* **103**, 2052–2086 (2013).
21. E. A. Hanushek, L. Woessmann, *J. Econ. Lit.* **46**, 607–668 (2008).
22. E. A. Hanushek, L. Woessmann, *J. Econ. Growth* **17**, 267–321 (2012).
23. H. B. Kim, S. Choi, B. Kim, C. Pop-Eleches, The role of education interventions in improving economic rationality, Version 1, CISER Data Archive (2018); https://doi.org/10.6077/80ph-y162.

MUSCLE DISEASE

Gene editing restores dystrophin expression in a canine model of Duchenne muscular dystrophy

Leonela Amoasii[1,2], John C. W. Hildyard[3], Hui Li[1], Efrain Sanchez-Ortiz[1], Alex Mireault[1], Daniel Caballero[1], Rachel Harron[3], Thaleia-Rengina Stathopoulou[4], Claire Massey[3], John M. Shelton[5], Rhonda Bassel-Duby[1], Richard J. Piercy[3], Eric N. Olson[1]*

Mutations in the gene encoding dystrophin, a protein that maintains muscle integrity and function, cause Duchenne muscular dystrophy (DMD). The deltaE50-MD dog model of DMD harbors a mutation corresponding to a mutational "hotspot" in the human *DMD* gene. We used adeno-associated viruses to deliver CRISPR gene editing components to four dogs and examined dystrophin protein expression 6 weeks after intramuscular delivery (n = 2) or 8 weeks after systemic delivery (n = 2). After systemic delivery in skeletal muscle, dystrophin was restored to levels ranging from 3 to 90% of normal, depending on muscle type. In cardiac muscle, dystrophin levels in the dog receiving the highest dose reached 92% of normal. The treated dogs also showed improved muscle histology. These large-animal data support the concept that, with further development, gene editing approaches may prove clinically useful for the treatment of DMD.

The persistent contraction and relaxation of cardiac and skeletal muscles necessitates mechanisms that maintain the integrity of muscle membranes (*1*, *2*). Dystrophin is a large scaffolding protein that supports muscle structure and function by linking the cytoskeleton with the sarcolemma of muscle tissue (*1*, *3*, *4*). Mutations in the dystrophin gene cause Duchenne muscular dystrophy (DMD), a disorder primarily affecting boys that is characterized by progressive muscle degeneration and atrophy, leading to premature death from cardiomyopathy and respiratory collapse (*5*). Thousands of mutations have been identified in the dystrophin gene, which spans ~2.5 megabases of DNA and contains 79 exons. Many of these mutations cluster into "hotspots," most commonly in a region that spans exons 45 to 50, typically placing exon 51 out of frame with preceding exons and preventing expression of functional dystrophin protein (*6*, *7*). Therapies that induce "skipping" of exon 51 restore the reading frame and in principle could benefit ~13% of DMD patients (*8*). An oligonucleotide that allows skipping of exon 51 can restore dystrophin expression to 0.22 to 0.32% of normal levels after 1 year of treatment and has been approved for DMD patients (*9–13*).

CRISPR-Cas9 gene editing can target DMD mutations and restore dystrophin expression in mice and muscle cells derived from human induced pluripotent stem cells (iPSCs) (*14–22*). An essential step toward clinical translation of gene editing as a therapeutic strategy for DMD is the demonstration of efficacy and safety of this approach in large mammals.

The mutation carried by the deltaE50-MD canine model of DMD leads to loss of exon 50 and moreover can be corrected by skipping of exon 51, making this a valuable model for translational studies. First identified as a naturally occurring, spontaneous mutation in Cavalier King Charles Spaniels (*23*) and now maintained on a beagle background, this model (in contrast to mice) exhibits many of the clinical and pathological

[1]Department of Molecular Biology, Hamon Center for Regenerative Science and Medicine, Sen. Paul D. Wellstone Muscular Dystrophy Cooperative Research Center, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390, USA. [2]Exonics Therapeutics, 75 Kneeland Street, Boston, MA 02111, USA. [3]Department of Clinical Science and Services, Comparative Neuromuscular Diseases Laboratory, Royal Veterinary College, London NW1 0TU, UK. [4]Section of Anaesthesia and Analgesia, Royal Veterinary College, London NW1 0TU, UK. [5]Department of Internal Medicine, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390, USA.
*Corresponding author. Email: eric.olson@utsouthwestern.edu
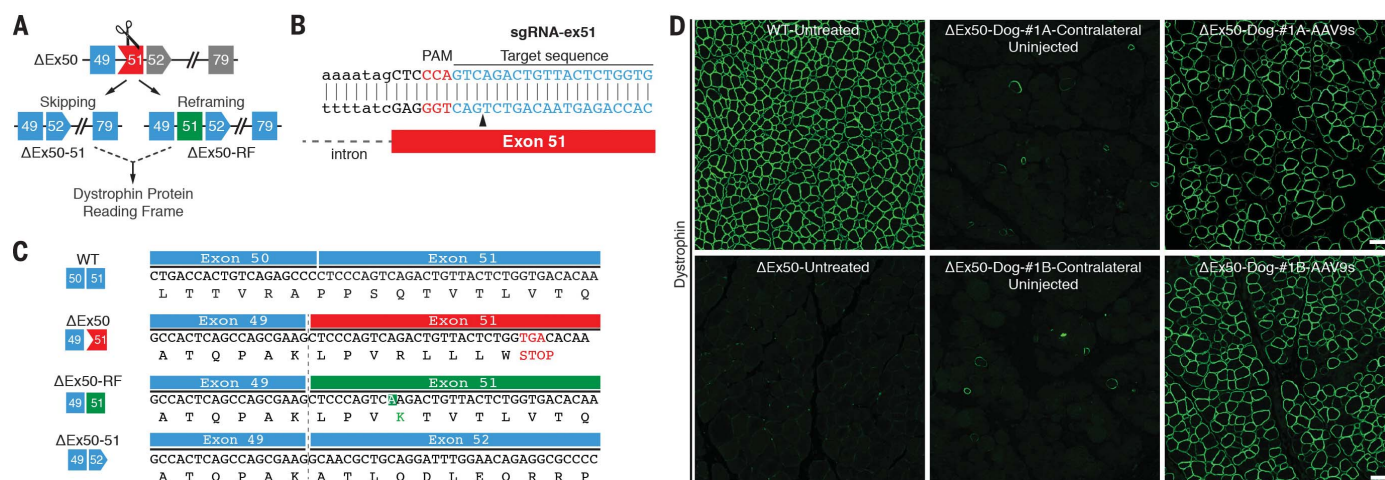
Fig. 1. Single-cut CRISPR editing of canine exon 50 in vivo and in vitro. (**A**) Scheme showing the CRISPR-Cas9–mediated genome editing approach to correct the reading frame in ΔEx50 dogs by reframing and skipping of exon 51. Gray exons are out of frame. (**B**) Illustration of sgRNA binding position and sequence for sgRNA-ex51. PAM sequence for sgRNA is indicated in red. Black arrowhead indicates the cleavage site. (**C**) Sequence of the RT-PCR products of the ΔEx50-51 lower band confirmed that exon 49 spliced directly to exon 52, excluding exon 51. Sequence of RT-PCR products of ΔEx50 reframed (ΔEx50-RF). (**D**) Cranial tibialis muscles of ΔEx50 dogs were injected with AAV9s encoding sgRNA-51 and Cas9 and analyzed 6 weeks later. Dystrophin immunohistochemistry staining of cranial tibialis muscle of wild-type (WT) dog untreated, ΔEx50 dog untreated, ΔEx50 dogs contralateral (uninjected) muscle, and ΔEx50 dogs injected with AAV9-Cas9 and AAV9-sgRNA-51 (referred to as ΔEx50-#1A-AAV9s and ΔEx50-#1B-AAV9s). Scale bar: 50 μm.

features of the human disease, such as muscle weakness, atrophy, and fibrosis (*24*).

To correct the dystrophin reading frame in the deltaE50-MD canine model (henceforth referred to as ΔEx50) (Fig. 1A), we used *Streptococcus pyogenes* Cas9 coupled with a single guide RNA (sgRNA) to target a region adjacent to the exon 51 splice acceptor site (referred to as sgRNA-51) (Fig. 1B). The sgRNA-51 corresponded to a highly conserved sequence that differs by only one nucleotide between the human and dog genomes (fig. S1A). Cas9 coupled with each of these sgRNA-51 sequences introduced a genomic cut only in DNA of the respective species (fig. S1B).

For the in vivo delivery of Cas9 and sgRNA-51 to skeletal muscle and heart tissue in dogs, we used recombinant adeno-associated virus serotype 9 (referred to as AAV9), which displays preferential tropism for these tissues (*25*, *26*). A muscle-specific creatine kinase (CK) regulatory cassette was used to drive expression of Cas9; three RNA polymerase III promoters (U6, H1, and 7SK) directed expression of the sgRNA, as described previously in mice (fig. S2) (*18*). AAV9-Cas9 and AAV9-sgRNA-51 were initially introduced into the cranial tibialis muscles of two 1-month-old dogs by intramuscular injection with 1.2×10^{13} AAV9 viral genomes (vg) of each virus. Muscles were analyzed 6 weeks after injection. In vivo targeting efficiency was estimated within muscle biopsy samples by reverse transcription–polymerase chain reaction (RT-PCR) with primers for sequences in exons 48 and 53, and genomic PCR amplification products spanning the target site were subjected to amplicon deep sequencing. The latter indicated that a mean of 9.96% of total reads contained changes at the targeted genomic site, including insertions, deletions, and substitutions (fig. S3). The most commonly identified mutations with a mean of 2.35% contained an adenosine (A) insertion immediately 3′ to the Cas9 genomic cutting site (Fig. 1C). The deletions identified with this method encompassed a highly predicted exonic splicing enhancer (ESE) site for exon 51 (*18*, *27*, *28*) (fig. S3A). However, this method does not identify larger deletions that might occur beyond the annealing sites of the primers used for PCR. Because these tissue samples contain a mixture of muscle and nonmuscle cells, the method likely underestimates the efficiency of gene editing within muscle cells.

Sequencing of RT-PCR products of RNA from muscle of ΔEx50 dogs injected intramuscularly with AAV9-Cas9 and AAV9-sgRNA-51 showed that deletion of exon 51 (ΔEx50-51) allowed splicing from exon 49 to 52, which restores the dystrophin open reading frame (fig. S3B). On gels, the PCR product with the A insertion was indistinguishable in size from non-edited cDNA products, so we performed deep-sequencing analysis to quantify its abundance compared to other small insertions. Deep sequencing of the upper band containing the non-edited cDNA product and reframed cDNA products indicated that a mean of 73.19% of total reads contained reframed cDNA products with an A insertion, a mean of

26.81% contained non-edited cDNA product, and the rest contained small deletions and insertions (fig. S3C). However, nonsense-mediated decay might affect the abundance of non-edited cDNA products. These data indicate that the two ΔEx50 dogs injected with AAV9-Cas9 and AAV9-sgRNA-51 had a high frequency of reframing events (with cDNA products containing an A insertion in the sequence of exon 51) and exon 51 skipping events resulting from deletion of the highly conserved ESE region.

To evaluate the specificity of our gene editing approach, we analyzed predicted off-target genomic sites for possible promiscuous editing. A total of three potential genome-wide off-target sites (OT1 to OT3) (fig. S4) were predicted in coding exons and four in noncoding regions (fig. S4) by the CRISPR design tool (http://crispr.mit.edu/). We performed deep sequencing at the top predicted off-target sites within protein-coding exons. None of these sites revealed notably more sequence alterations than the background analysis performed with other regions of the amplicons (fig. S5).

To evaluate dystrophin correction at the protein level, we performed histological analysis of AAV9-injected cranial tibialis muscles 6 weeks after AAV9 injection. Dystrophin immunohisto-

chemistry of muscle from ΔEx50 dogs injected with AAV9-Cas9 and AAV9-sgRNA-51 revealed widespread expression: The majority of fibers within the injected muscles expressed sarcolemmal dystrophin, albeit to varying levels (Fig. 1D). Western blot analysis confirmed the restoration of dystrophin expression in skeletal muscle (Fig. 2, A and B) to ~60% of wild-type (WT) levels. On average, 2% of WT levels of dystrophin were detected in the uninjected contralateral muscles, far more than could be attributed to rare revertant events, which typically represent fewer than 0.001% of fibers and are undetectable by Western blot in ΔEx50 muscle. We attribute expression in uninjected contralateral muscles to leakage of AAV9 into circulation. As assessed by hematoxylin and eosin (H&E) staining, the injected muscles appeared to be normalized relative to muscles of untreated animals, with fewer hypercontracted or necrotic fibers, reduced edema and fibrosis, and fewer regions of inflammatory cellular infiltration (Fig. 2C). Immunohistochemistry for developmental myosin heavy chain (dMHC), a marker of regenerating fibers, revealed a marked reduction in dMHC-positive fibers within injected muscles (fig. S6).

Dystrophin nucleates a series of proteins into the dystrophin-associated glycoprotein complex
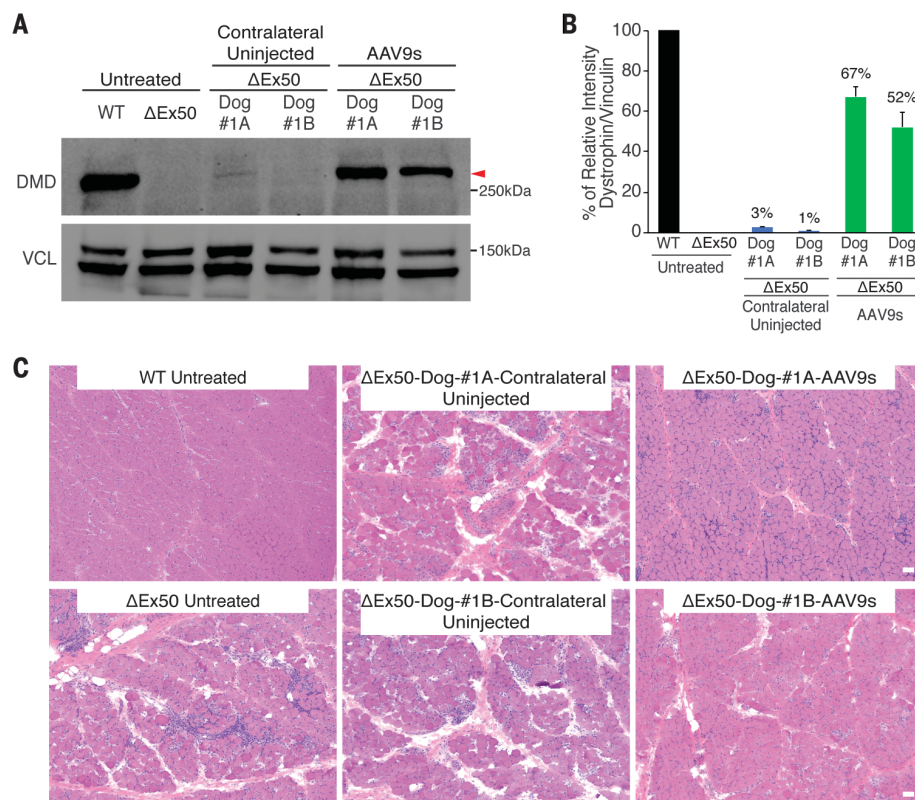
Fig. 2. Dystrophin correction after intramuscular delivery of AAV9-encoded gene editing components. (**A**) Western blot analysis of dystrophin (DMD) and vinculin (VCL) expression in cranial tibialis muscles 6 weeks after intramuscular injection in 2 dogs (#1A and #1B). (**B**) Quantification of dystrophin expression from blots after normalization to vinculin. (**C**) Histochemistry by H&E staining of cranial tibialis muscle from a WT dog, ΔEx50 dog untreated, ΔEx50 contralateral uninjected, and ΔEx50 dogs injected intramuscularly with AAV9-Cas9 and AAV9-sgRNA-51 (referred to as ΔEx50-Dog-#1A-AAV9s and ΔEx50-Dog-#1B-AAV9s). Scale bar: 50 μm.
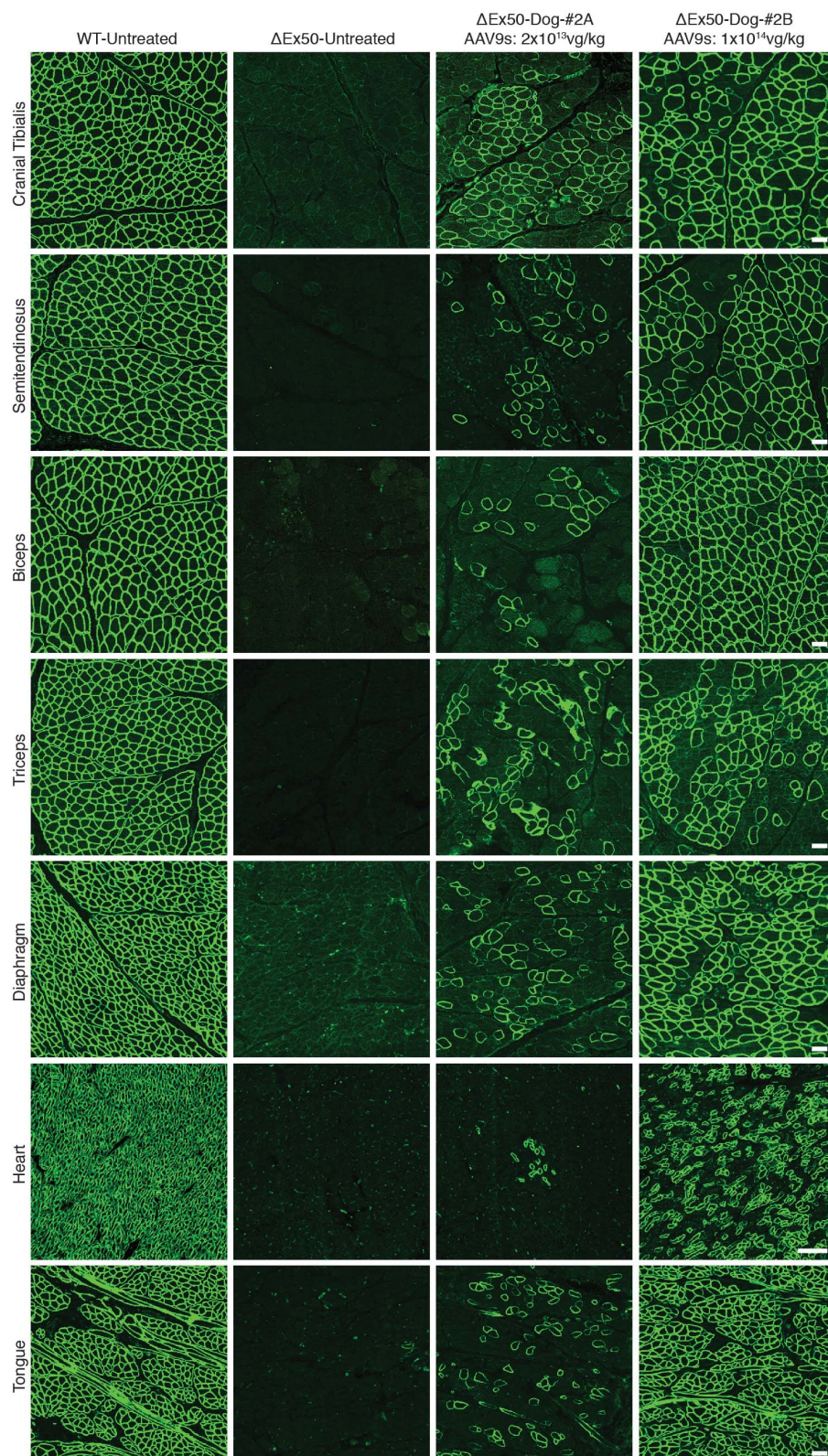
Fig. 3. Immunostaining of dystrophin after intravenous delivery of AAV9-encoded gene editing components. Dystrophin immunohistochemistry staining of cranial tibialis, semitendinosus, biceps, triceps, diaphragm, heart, and tongue muscles of WT dog, untreated ΔEx50 dog, and ΔEx50 dogs injected systemically with AAV9-Cas9 and AAV9-sgRNA-51 at 2×10^{13} vg/kg (total virus 4×10^{13} vg/kg, referred to as ΔEx50-Dog #2A-AAV9s) and 1×10^{14} vg/kg (total virus 2×10^{14} vg/kg, referred to as ΔEx50-Dog #2B-AAV9s) for each virus.

(DGC) to link the cytoskeleton and extracellular matrix (3, 4). In ΔEx50 mice, dogs, and DMD patients, these proteins are destabilized and do not localize to the subsarcolemmal region (4). Muscles injected with AAV9-Cas9 and AAV9-sgRNA-51 showed recovery of the DGC protein β-dystroglycan compared to contralateral uninjected muscles (fig. S7). We conclude that—at least in a short time frame of 6 weeks—single-cut genomic editing using AAV9-Cas9 and AAV9-sgRNA-51 can efficiently restore dystrophin expression and assembly of the DGC in dystrophic muscles. Immunohistochemistry using canine-specific CD4 and CD8 T cell markers (fig. S8) showed no evidence of enhanced mononuclear cellular infiltration or relevant hematological abnormalities in the treated animals compared to untreated controls or reference ranges (fig. S9).

Based on the high dystrophin-correction efficiency observed after intramuscular injection of AAV9-Cas9 and AAV9-sgRNA-51, we tested for rescue of dystrophin expression in two ΔEx50 dogs after systemic delivery of gene editing components. The dogs at 1 month of age were injected intravenously with the viruses and analyzed 8 weeks later. We tested two doses (2×10^{13} vg/kg and 1×10^{14} vg/kg) of each of the two viruses (AAV9-Cas9 and AAV9-sgRNA-51). To avoid a possible immune reaction, we included a transient regimen of immune suppression with the high dose. Systemic delivery of 2×10^{13} vg/kg of each virus (total virus 4×10^{13} vg/kg) in ΔEx50-Dog-#2A resulted in expression of virus in peripheral skeletal muscle samples and, to a lower extent, in heart samples, as shown by quantitative PCR (qPCR) analysis (fig. S10A). The delivery of 1×10^{14} vg/kg of each virus (total virus 2×10^{14} vg/kg) in ΔEx50-Dog-#2B (via infusion) allowed more widespread expression of viral constructs in the peripheral skeletal muscle samples and in heart samples (fig. S10B). Systemic delivery of AAV9-Cas9 and AAV9-sgRNA-51 led to dystrophin expression in a broad range of muscles, including the heart, in gene-edited ΔEx50 dogs at 8 weeks after injection—and to a markedly greater extent than that achieved with the lower dose (Fig. 3).

To investigate the proportions of various insertions and deletions (indels) generated by systemic delivery of AAV9-Cas9 and AAV9-sgRNA-51, we performed amplicon deep-sequencing analysis of the genomic DNA from heart, triceps, and biceps muscles. The genomic deep-sequencing analysis revealed an increased percentage of reads containing changes at the targeted genomic site, especially of the 1A insertion mutation in the samples from Dog-#2B compared to Dog-#2A (fig. S11).

Additionally, we performed tracking indels by decomposition (TIDE) (30) analysis at the genomic and cDNA levels, which showed an increase in numbers of indels in the samples from Dog-#2B compared to the samples from Dog-#2A (fig. S12). Testes analysis and Western blot analysis showed no activity of Cas9 and confirmed muscle-specific expression of gene editing machinery (fig. S13). Western blot analysis confirmed

the restoration of dystrophin expression in skeletal muscle (Fig. 4, A and B) to ~50, 20, and 3% of WT levels for the cranial tibialis, triceps, and biceps, respectively, after systemic delivery of 2×10^{13} vg/kg of each virus (total virus 4×10^{13} vg/kg). For Dog-#2B, which received 1×10^{14} vg/kg of each virus (total virus 2×10^{14} vg/kg), Western blot analysis showed restoration of dystrophin expression (Fig. 4, C and D, and fig. S14) to ~70, 25, 64, 58, 92, and 5% of WT levels for the cranial tibialis, triceps, biceps, diaphragm, heart, and tongue muscles, respectively. Similar to what was seen after intramuscular injection, muscles appeared normalized via H&E staining (Fig. 4E). Immunostaining of muscle sections from treated ΔEx50 dogs also showed recovery of β-dystroglycan expression (fig. S15) and widespread reduction in dMHC (fig. S16).

To determine hematological and biochemical parameters of the treated dogs compared to the controls, we collected blood samples the day before injection and then at 1, 2, 4, 6, and 8 weeks after injection. The blood samples from all four dogs [healthy untreated, ΔEx50 untreated, and ΔEx50 dogs receiving 2×10^{13} vg/kg of each virus (total virus 4×10^{13} vg/kg) and 1×10^{14} vg/kg of each virus (total virus 2×10^{14} vg/kg)] before and after injection were unremarkable (fig. S17). Hematology counts, serum electrolytes, and kidney and liver function parameters remained within the normal ranges in all dogs. Additionally, blood samples were collected weekly for CK assessment: We observed a modest decline in serum CK activity in Dog-#2B treated with 1×10^{14} vg/kg of each virus (total virus 2×10^{14} vg/kg) (fig. S18) compared to the ΔEx50 untreated dog.

To evaluate the targeting efficiency of a human DMD mutation, we used a DMD iPSC line carrying a deletion from exon 48 to 50. Deletion of exons 48 to 50 leads to a frameshift mutation and appearance of a premature stop codon in exon 51. To correct the dystrophin reading frame, we introduced two concentrations of Cas9 and sgRNA-51 (26 ng/μl, referred to as high, and 13 ng/μl, referred to as low). Indel analysis showed 55.8 and 31.9% of indels for the high and low concentrations, respectively (fig. S19). Genomic deep-sequencing analysis revealed that 27.94% of mutations contained a single A insertion 3′ to the protospacer adjacent motif (PAM) sequence for the high concentration condition and 19.03% for the low concentration condition (fig. S20A), as observed in mouse and dog cells with a similar sgRNA directed against the same genomic locus. DMD iPSCs treated with Cas9 and sgRNA-51 and induced to form cardiomyocytes (iCMs) showed restoration of dystrophin immunostaining (fig. S20B) and expression of dystrophin protein to 67 to 100% of the levels of WT cardiomyocytes, as measured by Western blot (fig. S20, C and D).

It has been estimated that even 15% of normal levels of dystrophin would provide substantial therapeutic benefits for DMD patients (30–32). Our results demonstrate the efficacy of single-cut genome editing for restoration of dystrophin expression in a large-animal model of DMD, reaching up to ~80% of WT levels in some muscles
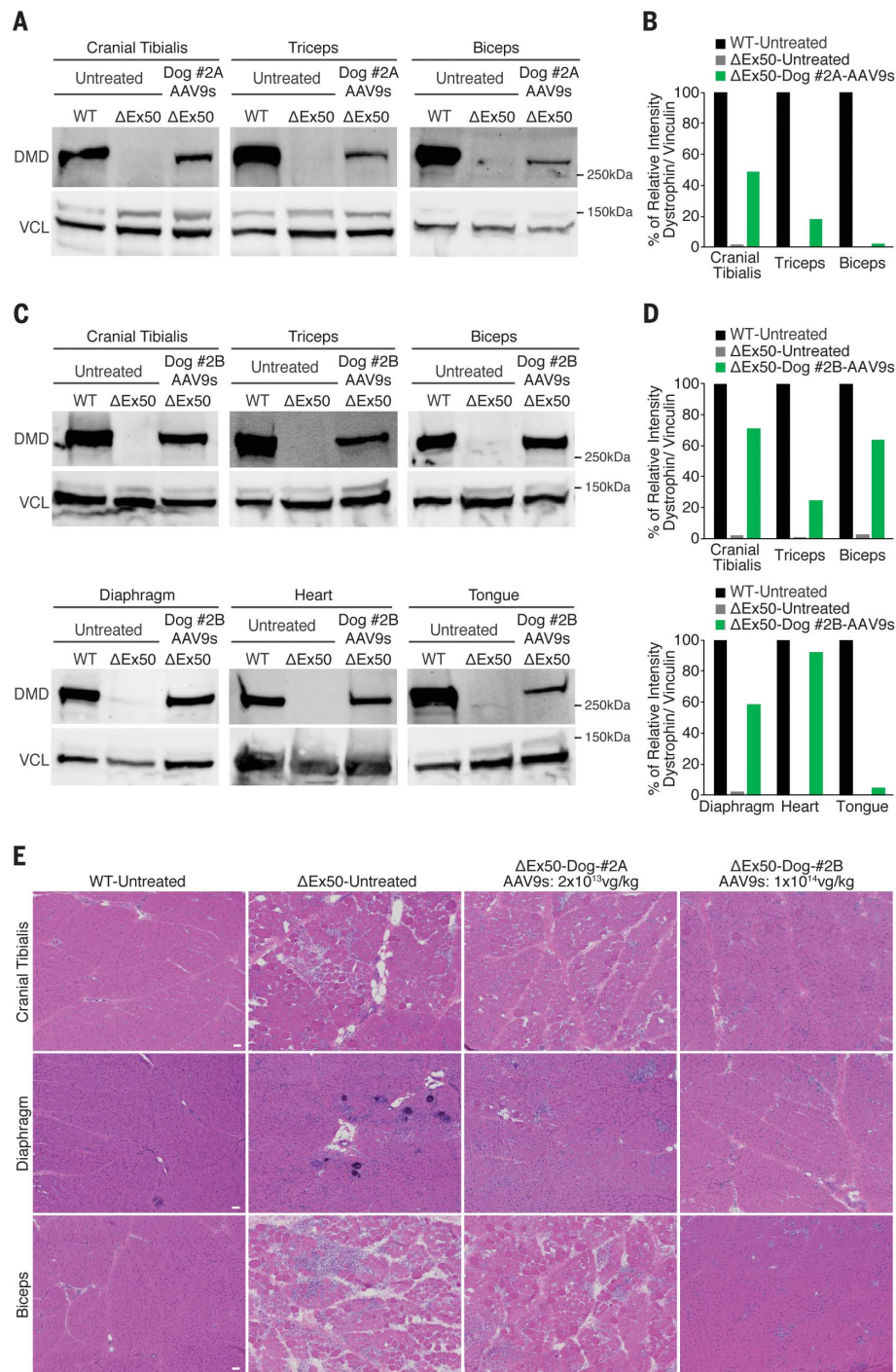
Fig. 4. Western blot of dystrophin and muscle histology after intravenous delivery of AAV9-encoded gene editing components. (**A**) Western blot analysis of dystrophin (DMD) and vinculin (VCL) of cranial tibialis, triceps, and biceps muscles of WT, untreated ΔEx50, and ΔEx50 injected with AAV9-Cas9 and AAV9-sgRNA-51 at 2×10^{13} vg/kg for each virus (total virus 4×10^{13} vg/kg, referred to as ΔEx50-Dog #2A-AAV9s). (**B**) Quantification of dystrophin expression from blots after normalization to vinculin. (**C**) Western blot analysis of DMD and VCL of cranial tibialis, triceps, biceps, diaphragm, heart, and tongue muscles of WT, untreated ΔEx50, and ΔEx50 injected with AAV9-Cas9 and AAV9-sgRNA-51 at 1×10^{14} vg/kg (total virus 2×10^{14} vg/kg, referred to as ΔEx50-Dog #2B-AAV9s). (**D**) Quantification of dystrophin expression from blots after normalization to vinculin. (**E**) H&E staining of cranial tibialis, diaphragm, and biceps muscles of WT, untreated ΔEx50, and ΔEx50 injected with AAV9-Cas9 and AAV9-sgRNA-51 at 2×10^{13} vg/kg for each virus (total virus 4×10^{13} vg/kg) and 1×10^{14} vg/kg for each virus (total virus 2×10^{14} vg/kg). Scale bar: 50 μm.

after 8 weeks. Longer-term studies are required to establish whether the expression of dystrophin and the maintenance of muscle integrity that we observed are sustained.

This study, while encouraging, is preliminary and has several limitations, including the relatively small number of animals analyzed and the short duration of the analysis (6 to 8 weeks). The possibility of off-target effects of in vivo gene editing is a further potential safety concern. Although our initial deep-sequencing analysis of the top predicted off-target sites revealed no specific off-target gene editing above background levels in treated animals, it will be important to further assess possible off-target mutagenesis in longer-term studies with greater numbers of animals. Recent studies reported large deletions and complex genomic rearrangements at target sites of CRISPR-Cas9 in mouse embryonic cells, hematopoietic progenitors, and human immortalized epithelial cells (*33*). However, these cells are highly proliferative, are more genomically unstable in culture, and use different DNA repair pathways than somatic postmitotic cells (such as muscle and heart cells) (*34*). Future studies will be required to investigate the long-term genomic stability of gene-edited muscle tissues in vivo. Another potential concern with CRISPR-Cas9–mediated gene editing in vivo is immunogenicity of Cas9, particularly as expression persists after initial treatment. In this short-term study, we did not observe CD4- or CD8-positive cell infiltration of treated muscle, but longer-term studies with more sensitive assays must be performed. Additionally, although production of large quantities of AAV9 poses a challenge, doses of 2×10^{14} vg/kg have been successfully used in human gene therapy trials (*35*).

Finally, although gene editing and exon-skipping oligonucleotides can both restore production of internally deleted dystrophin proteins, similar to those expressed in Becker muscular dystrophy, there are key distinctions between these approaches. Most notably, CRISPR gene editing may be permanent and not require redelivery, whereas oligonucleotides require continuous treatment. A corollary of this is that CRISPR treatment may be difficult to terminate if safety concerns arise.

REFERENCES AND NOTES

1. K. F. O'Brien, L. M. Kunkel, *Mol. Genet. Metab.* **74**, 75–88 (2001).
2. S. Guiraud et al., *Annu. Rev. Genomics Hum. Genet.* **16**, 281–308 (2015).
3. K. P. Campbell, S. D. Kahl, *Nature* **338**, 259–262 (1989).
4. J. M. Ervasti, K. Ohlendieck, S. D. Kahl, M. G. Gaver, K. P. Campbell, *Nature* **345**, 315–319 (1990).
5. F. Muntoni, S. Torelli, A. Ferlini, *Lancet Neurol.* **2**, 731–740 (2003).
6. C. L. Bladen et al., *Hum. Mutat.* **36**, 395–402 (2015).
7. Y. Li et al., *J. Int. Med. Res.* **44**, 99–108 (2016).
8. A. Aartsma-Rus et al., *Hum. Mutat.* **30**, 293–299 (2009).
9. Y. Shimizu-Motohashi, S. Miyatake, H. Komaki, S. Takeda, Y. Aoki, *Am. J. Transl. Res.* **8**, 2471–2489 (2016).
10. C. A. Stein, *Mol. Ther.* **24**, 1884–1885 (2016).
11. A. Aartsma-Rus, A. M. Krieg, *Nucleic Acid Ther.* **27**, 1–3 (2017).
12. J. J. Dowling, *Nat. Rev. Neurol.* **12**, 675–676 (2016).
13. J. R. Mendell et al., *Ann. Neurol.* **79**, 257–271 (2016).
14. C. Long et al., *Science* **351**, 400–403 (2016).
15. M. Tabebordbar et al., *Science* **351**, 407–411 (2016).
16. C. E. Nelson et al., *Science* **351**, 403–407 (2016).
17. C. Long et al., *Sci. Adv.* **4**, p9004 (2018).
18. L. Amoasii et al., *Sci. Transl. Med.* **9**, eaan8081 (2017).
19. N. E. Bengtsson et al., *Nat. Commun.* **8**, 14454 (2017).
20. J. P. Tremblay, J.-P. Iyombe-Engembe, B. Duchêne, D. L. Ouellet, *Mol. Ther.* **24**, 1888–1889 (2016).
21. C. S. Young et al., *Cell Stem Cell* **18**, 533–540 (2016).
22. D. G. Ousterout et al., *Nat. Commun.* **6**, 6244 (2015).
23. G. L. Walmsley et al., *PLOS ONE* **5**, e8647 (2010).
24. J. Hildyard et al., *Neuromuscul. Disord.* **28**, S18 (2018).
25. Y. Yue et al., *Hum. Mol. Genet.* **24**, 5880–5890 (2015).
26. C. Zincarelli, S. Soltys, G. Rengo, J. E. Rabinowitz, *Mol. Ther.* **16**, 1073–1080 (2008).
27. P. J. Smith et al., *Hum. Mol. Genet.* **15**, 2490–2508 (2006).
28. L. Cartegni, J. Wang, Z. Zhu, M. Q. Zhang, A. R. Krainer, *Nucleic Acids Res.* **31**, 3568–3571 (2003).
29. E. K. Brinkman, T. Chen, M. Amendola, B. van Steensel, *Nucleic Acids Res.* **42**, e168 (2014).
30. C. Godfrey et al., *Hum. Mol. Genet.* **24**, 4225–4237 (2015).
31. M. van Putten et al., *J. Mol. Cell. Cardiol.* **69**, 17–23 (2014).
32. A. H. Beggs et al., *Am. J. Hum. Genet.* **49**, 54–67 (1991).
33. M. Kosicki, K. Tomberg, A. Bradley, *Nat. Biotechnol.* **36**, 765–771 (2018).
34. E. D. Tichy et al., *Stem Cells Dev.* **19**, 1699–1711 (2010).
35. J. R. Mendell et al., *N. Engl. J. Med.* **377**, 1713–1722 (2017).

CANCER

Reprogramming normal human epithelial tissues to a common, lethal neuroendocrine cancer lineage

Jung Wook Park[1], John K. Lee[2], Katherine M. Sheu[3], Liang Wang[1], Nikolas G. Balanis[3], Kim Nguyen[4], Bryan A. Smith[1], Chen Cheng[5], Brandon L. Tsai[1], Donghui Cheng[1], Jiaoti Huang[6], Siavash K. Kurdistani[5,7,8,9], Thomas G. Graeber[3,7,8,9,10]*, Owen N. Witte[1,3,7,8,9]*

The use of potent therapies inhibiting critical oncogenic pathways active in epithelial cancers has led to multiple resistance mechanisms, including the development of highly aggressive, small cell neuroendocrine carcinoma (SCNC). SCNC patients have a dismal prognosis due in part to a limited understanding of the molecular mechanisms driving this malignancy and the lack of effective treatments. Here, we demonstrate that a common set of defined oncogenic drivers reproducibly reprograms normal human prostate and lung epithelial cells to small cell prostate cancer (SCPC) and small cell lung cancer (SCLC), respectively. We identify shared active transcription factor binding regions in the reprogrammed prostate and lung SCNCs by integrative analyses of epigenetic and transcriptional landscapes. These results suggest that neuroendocrine cancers arising from distinct epithelial tissues may share common vulnerabilities that could be exploited for the development of drugs targeting SCNCs.

Human cancers originating in different organs share commonalities in cancer phenotypes and molecular features (1). Small cell neuroendocrine carcinoma (SCNC) is recognized by its distinctive histological features and can arise from almost all epithelial organs, including the prostate and lung (2, 3). These cancers show notable lineage plasticity and acquire therapeutic resistance by converting from an epithelial to a neuroendocrine cancer phenotype (4–8). Large-scale analyses of transcriptome data from a variety of cancer types have provided substantial evidence (9–11) to support a phenotypic convergence to SCNC during cancer progression. The underlying molecular mechanisms are not fully understood.

To explore whether distinct human epithelial cell types can be transformed into SCNC by shared oncogenic drivers, we used a human tissue transformation assay (12). Small cell prostate cancer (SCPC) is a type of neuroendocrine prostate cancer (NEPC), a class of malignancies that includes the extremely rare "large cell prostate carcinoma," whose exact definition is still emerging (13). Overexpression of c-Myc or N-Myc in combination with myristoylated AKT1 (myrAKT1) (a partial mimic of *PTEN* loss) drives normal human prostate epithelial cells to poorly differentiated prostate adenocarcinoma (PrAd) or SCPC, respectively (12, 14). To identify additional oncogenic drivers of SCPC,

we investigated the functional effects of dominant negative p53 (TP53DN) (P), myrAKT1 (A), *RB1*–short hairpin RNA (shRNA) (R), c-Myc (C), and BCL2 (B) on the genesis of SCPC on the basis of their recurrent genetic alterations in prostate cancers, including SCPC (5, 7, 15, 16). These five genetic factors are hereafter referred to as PARCB. Primary basal epithelial cells were isolated from the prostates of eight human donors and were lentivirally transduced with the PARCB factors, briefly cultured in an organoid system (Fig. 1A), and transplanted into immunodeficient NOD/SCID-IL2R-γc-KO mice (17), where they formed tumors that displayed green, red, and yellow fluorescent protein (GFP, RFP, and YFP) expression from the three lentiviruses (Fig. 1B). PARCB gene expression and the human cellular origin of the tumors were confirmed by immunostaining (fig. S1). All PARCB tumors derived from basal cells demonstrated histological features of human SCPC, including a high nuclear-to-cytoplasmic ratio, frequent mitotic and apoptotic figures, and uniform expression of neuroendocrine differentiation (NED) markers (Fig. 1, C and D).

We next defined the genetic factors required to initiate SCPC from the PARCB combination by leave-one-out analysis. No tumors developed in the absence of either c-Myc or myrAKT1. In the absence of BCL2, the PARC combination still produced tumors with histologic characteristics

and NED marker expression consistent with SCPC (Fig. 1D and fig. S2A), albeit with reduced efficiency (fig. S3). In contrast to BCL2, *RB1*-shRNA and TP53DN together or individually were indispensable for SCPC development. Tumors arising from these conditions—PACB (without *RB1*-shRNA), ARCB (without TP53DN), and ACB (without TP53DN and *RB1*-shRNA)—displayed histological features of poorly differentiated PrAd. PACB and ARCB tumors displayed only focal expression of NED markers (Fig. 1D). Thus, loss of *RB1* and inactivation of p53 are required to convert an epithelial lineage into a neuroendocrine lineage in SCPC development during human prostate epithelial transformation.

To further investigate the molecular contributions of the PARCB genetic factors to SCPC, we established tumor cell lines from fluorescence-activated cell sorting (FACS)–purified cells of the ACB, PACB, ARCB, and PARCB tumors (Fig. 2A). Immunoblot analysis confirmed the expression of the respective genetic factors in the newly generated cell lines (fig. S2B). We then performed two downstream global analyses: mRNA sequencing (RNA-seq) and assay for transposase-accessible chromatin sequencing (ATAC-seq). Currently, gene expression datasets specific for SCPC are lacking. We used the largest available RNA-seq dataset of NEPC and PrAd patient samples (8) in our study. We simplified the nomenclature of NEPC as SCPC to prevent confusion when alternating between epithelial tissue types.

Our RNA-seq data revealed that the PARCB cell lines have transcriptomes that are distinct from those characterizing ACB, PACB, and ARCB lines (fig. S4A), supporting the histologic and molecular differences described above. The PARCB cell lines exhibited enriched expression of genes that are up-regulated in clinical SCPC specimens relative to PrAd samples (8, 18), whereas the ACB, PACB, and ARCB lines did not (fig. S5A). The PARCB lines were also highly similar to human SCPC (fig. S5B) on the basis of a published SCPC gene expression signature (7).

Global transcriptome analysis revealed that the PARCB cell lines exhibited strong transcriptional similarity to SCPC patient samples, whereas the other engineered cell lines clustered with patient-derived PrAd cell lines (Fig. 2B). They did not express detectable levels of androgen receptor (AR) and exhibited the lowest level of AR signaling activity when compared with clinical samples (fig. S6). The PARCB cell lines also exhibited NED markers in vivo and in vitro (fig. S7).

Open- or closed-chromatin regions can be indicative of transcriptional regulatory elements and serve as predictors of gene transcription activity. We measured genome-wide chromatin

[1]Department of Microbiology, Immunology, and Molecular Genetics, University of California–Los Angeles, Los Angeles, CA 90095, USA. [2]Division of Hematology and Oncology, Department of Medicine, University of California–Los Angeles, Los Angeles, CA 90095, USA. [3]Department of Molecular and Medical Pharmacology, University of California–Los Angeles, Los Angeles, CA 90095, USA. [4]Department of Ecology and Evolutionary Biology, University of California–Los Angeles, Los Angeles, CA 90095, USA. [5]Department of Biological Chemistry, University of California–Los Angeles, Los Angeles, CA 90095, USA. [6]Department of Pathology, School of Medicine, Duke University, Durham, NC 27710, USA. [7]Molecular Biology Institute, University of California–Los Angeles, Los Angeles, CA 90095, USA. [8]Jonsson Comprehensive Cancer Center, University of California–Los Angeles, Los Angeles, CA 90095, USA. [9]Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California–Los Angeles, Los Angeles, CA 90095, USA. [10]Crump Institute for Molecular Imaging, University of California–Los Angeles, Los Angeles, CA 90095, USA.
*Corresponding author. Email: owenwitte@mednet.ucla.edu (O.N.W.); tgraeber@mednet.ucla.edu (T.G.G.)
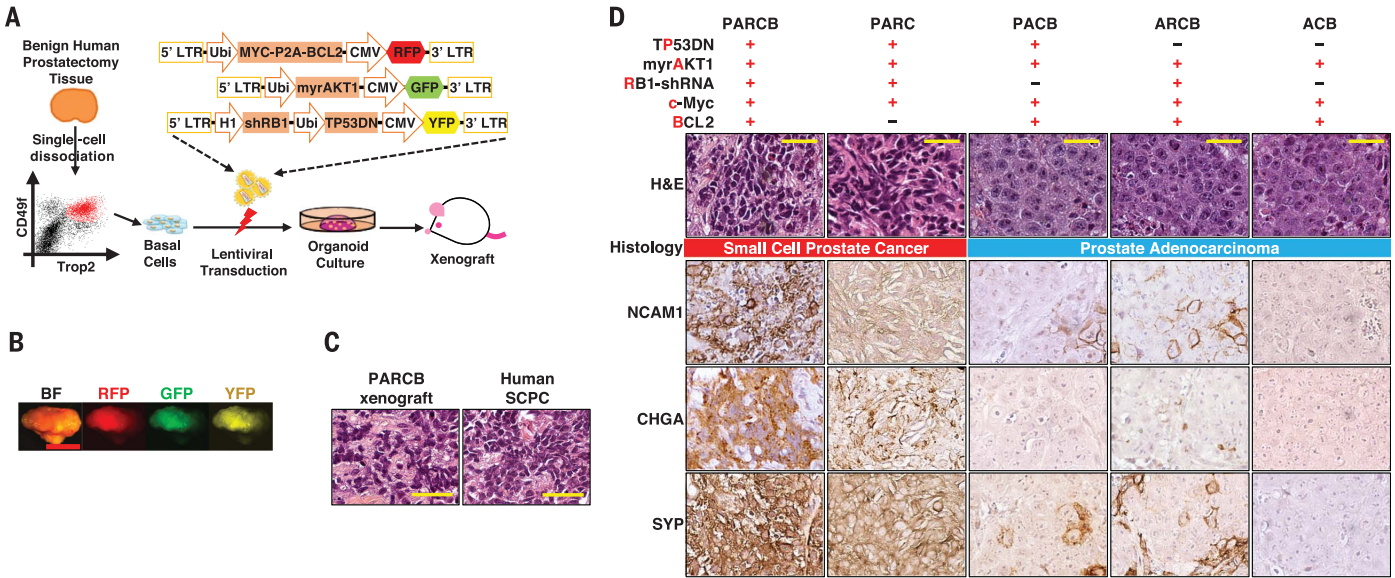
Fig. 1. Defined genes drive human prostate epithelial cells to the SCPC phenotype. (**A**) Schematic of human prostate transformation assay. LTR, long terminal repeats; CMV, cytomegalovirus promoter; Ubi, ubiquitin promoter. (**B**) Representative image of PARCB xenografts displaying GFP, RFP, and YFP expression. BF, bright field. Scale bar, 1 cm. (**C**) Hematoxylin and eosin (H&E)–stained images of PARCB grafts and human SCPC. Scale bars, 50 μm. (**D**) H&E staining and immunohistochemistry (IHC) images with antibodies against the indicated proteins in genetically engineered tumor models. Scale bars, 50 μm.
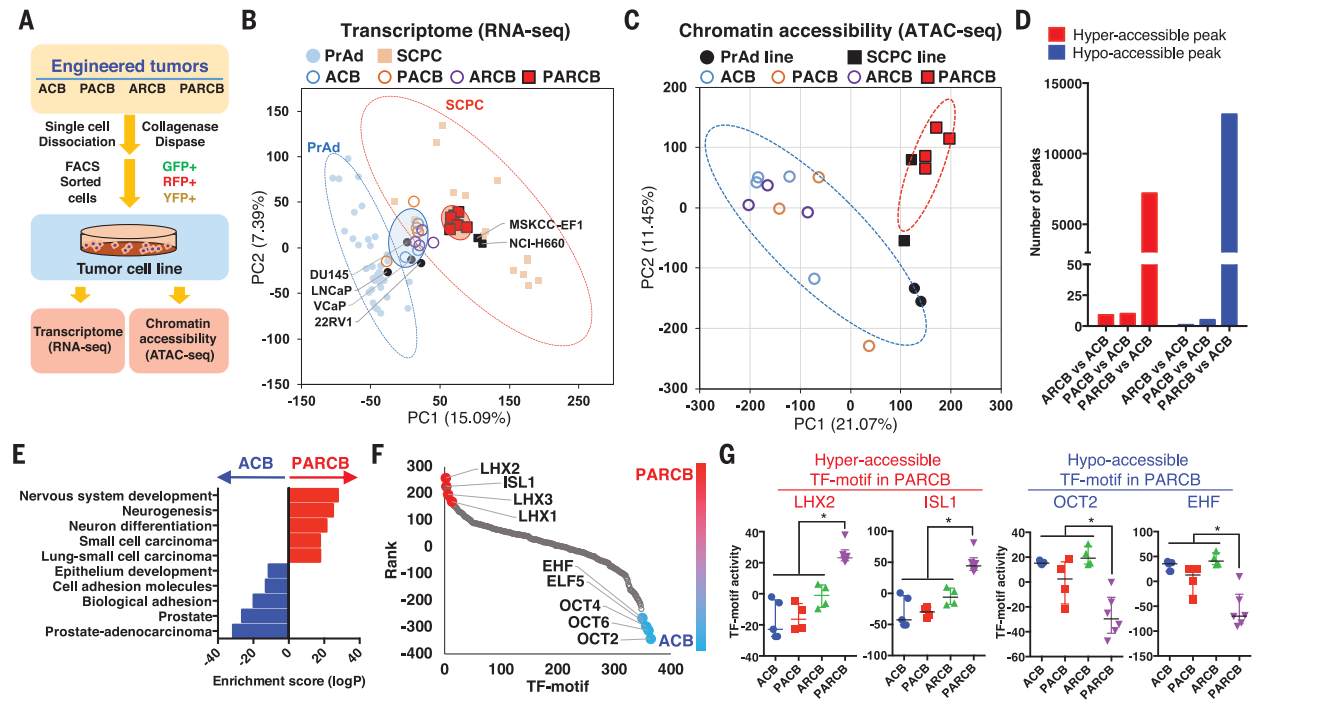
Fig. 2. Inactivation of both p53 and RB is required to reprogram transcriptional profiles and chromatin accessibility landscapes of normal prostate epithelial cells to human SCPC. (**A**) Schematic for establishment of tumor cell lines with GFP-, RFP-, and YFP-positive purified xenograft cells. (**B**) Partial least-squares regression analysis (PLSR) separates PrAd and SCPC specimens in the RNA-seq dataset from Beltran et al. (7). RNA-seq data for engineered tumor lines and patient-derived prostate cancer cell lines were projected onto the PLSR plot. (**C**) Principal components analysis (PCA) of ATAC-seq data from engineered cell lines with PrAd and SCPC lines. Probability ellipse = 95% confidence to group the samples. (**D**) Hyper- or hypo-accessible peaks in comparisons between engineered tumor lines. (**E**) Selected gene sets enriched in hyper- or hypo-accessible peaks in the comparison between PARCB and ACB lines. (**F**) TF binding motifs identified by HOMER (34) motif analysis were plotted by ranks generated from their associated differential adjusted P values. (**G**) Transcriptional activities of the TF motifs were measured by gene signature scores (see materials and methods). Medians with interquartile ranges are shown. $*P < 0.05$ [one-way analysis of variance (ANOVA)].
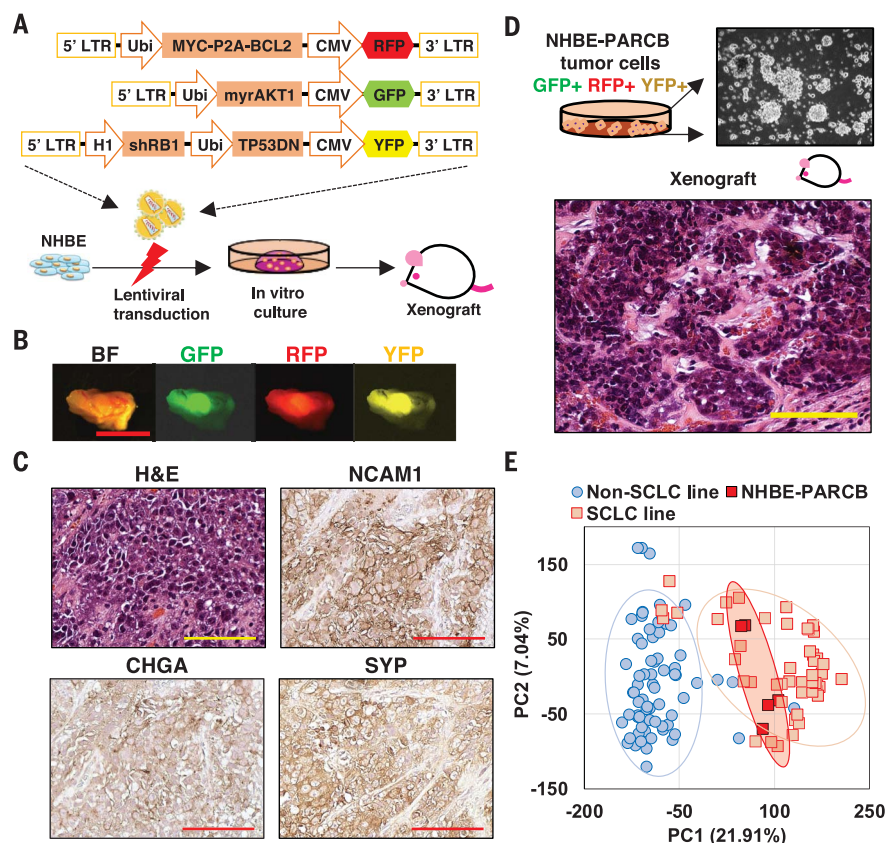
Fig. 3. The same genetic drivers for SCPC can initiate SCLC from human normal lung epithelial cells. (**A**) Schematic of NHBE cell transformation. (**B**) Representative tumor image displaying the lentivirally colinked markers GFP, RFP, and YFP. Scale bar, 1 cm. (**C**) H&E and IHC images against NED markers in transformed NHBE xenografts. Scale bars, 100 μm. (**D**) Representative morphologic image of NHBE-PARCB cell lines in two-dimensional culture and H&E image of xenografts derived from NHBE-PARCB cell lines. Scale bar, 100 μm. (**E**) PCA of NHBE-PARCB cell lines with SCLC and non-SCLC cell lines.

accessibility and its association with transcriptional programs by ATAC-seq. PARCB and patient-derived SCPC lines exhibited a distinct chromatin accessibility status compared with the ACB, PACB, or ARCB lines (Fig. 2C and fig. S4B). Dual inactivation of p53 and RB in PARCB lines induced marked changes in chromatin accessibility compared with that in ACB lines (Fig. 2D). However, single inactivation of p53 or RB alone in PACB or ARCB lines, respectively, did not alter chromatin accessibility compared with that in ACB lines (Fig. 2D).

Chromatin regions in PARCB lines that were hyperaccessible compared with those in ACB lines were highly enriched for genes associated with neuronal differentiation and small cell lung cancer (SCLC), whereas hypo-accessible regions in PARCB lines were enriched for epithelial development– and PrAd-associated genes (Fig. 2E). We conclude that concomitant p53 and RB disruption is required and may synergize to promote lineage plasticity during human prostate epithelial transformation by modulating transcriptional and epigenetic programs, supporting previous findings from mouse models (8, 19).

To further characterize p53 and RB inactivation–induced changes in chromatin, we performed transcription factor (TF) binding motif enrichment analysis with the differentially accessible peaks in PARCB lines compared with ACB lines. We found that LHX family TF motifs (LHX1, LHX2, and LHX3) and ISL1 were the most accessible regions in PARCB lines, whereas OCT family TF motifs (OCT2 and OCT6) and ETS family TF motifs (EHF and ELF5) were less accessible in PARCB than in ACB lines (Fig. 2F and table S1). Although the biological functions of these TFs are unexplored in prostate cancer, we were able to assess the transcriptional activity of these TF motifs altered by concomitant p53 and RB disruption. Our integrated analysis with the matched RNA-seq data revealed that enriched accessibility of LHX2-, ISL1-, OCT2-, or EHF-DNA binding motifs was mirrored by the expression of the downstream target genes (Fig. 2G).

We next investigated how similar the PARCB lines might be to SCNCs from other organs. We performed a gene list enrichment analysis (20) of 938 transcriptionally profiled cancer cell lines by using genes up-regulated in the PARCB lines

(table S2). This analysis demonstrated that the PARCB cell lines were transcriptionally most similar to SCLC cell lines and the NCI-H660 SCPC cell line (fig. S8 and table S3). To further evaluate the similarity of SCPC and SCLC, we projected gene expression data from human lung cancer samples [from (21) and The Cancer Genome Atlas (TCGA) (22)] onto our prostate cancer clustering plot in Fig. 2B. All clinical SCLC samples tightly clustered with the SCPC samples, indicating that prostate and lung SCNCs are both histologically and transcriptionally similar (fig. S9).

Given the similarity of the PARCB prostate cancer models to SCPC and SCLC, we next asked whether the PARCB factors could initiate SCLC from human lung epithelial cells. In a fashion similar to our transformation of primary human prostate cells, we transformed primary normal human bronchial epithelial (NHBE) cells with the PARCB lentiviruses (Fig. 3A). Xenografts (NHBE-PARCB) expressed GFP, RFP, and YFP (Fig. 3B) and the PARCB factors (fig. S10A). NHBE-PARCB tumors showed typical histological features and NED marker expression characteristic of clinical SCLC (Fig. 3C). Removing one gene from the PARCB combination did not result in tumor growth. Similar to clinical SCLC (23), all of the tumors expressed the proliferation marker Ki67 and none expressed the basal-squamous epithelial cell markers p63 and cytokeratin 14 (fig. S10B).

To interrogate the convergent evolutionary pattern of SCNCs, we established five NHBE-PARCB tumor cell lines for downstream multi-omics analyses. Like the prostate PARCB lines, the NHBE-PARCB lines propagated in suspension, demonstrated tumorigenic potential in vivo (Fig. 3D), and exhibited transcriptional similarity to patient-derived SCLC cell lines [from the Cancer Cell Line Encyclopedia database (24)] (Fig. 3E). Comparative genomic hybridization and whole-exome sequencing analyses identified no recurrent genetic alterations between prostate and lung PARCB models, indicating that the defined five genetic factors are likely sufficient to drive human SCNC (fig. S11 and tables S4 and S5).

We found that normal prostate basal epithelial cells and NHBE cells displayed distinct transcriptomes. In contrast, the prostate PARCB and NHBE-PARCB cell lines tightly clustered together, indicating a shared gene expression profile associated with reprogramming by the PARCB factors irrespective of the tissue of origin (Fig. 4A). We then examined transcriptome data from SCPC and SCLC patient biopsy specimens (7, 21) and compared these with their respective adjacent normal tissues from TCGA. These results support the finding that SCPC and SCLC are transcriptionally convergent relative to the normal epithelial cells from which they originated (Fig. 4B) and demonstrate that a similar convergence occurs in patients.

Both transcriptional and epigenetic regulatory networks dictate cell lineage decisions (25, 26). We investigated the nature of the global chromatin states involved in the convergent transformation to SCNC by ATAC-seq. PARCB, SCPC,
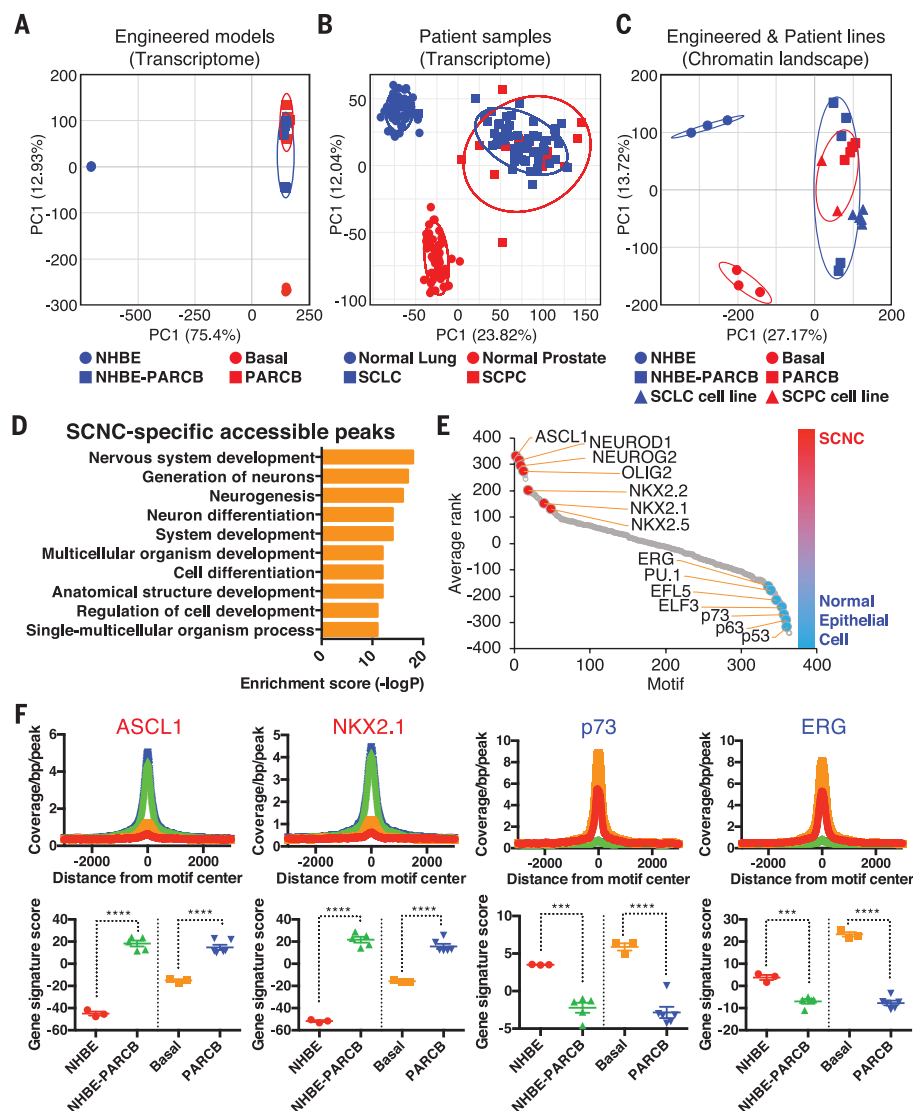
Fig. 4. Convergent transcriptional and chromatin accessibility landscapes of SCPC and SCLC identify shared TF motifs. (**A**) PCA of RNA-seq data from normal lung and prostate epithelial cells (NHBE and basal) and engineered lung and prostate PARCB cells. (**B**) PCA of RNA-seq data from human samples: SCLC and SCPC specimens and adjacent normal lung and prostate tissues. (**C**) PCA of ATAC-seq data from NHBE, basal, NHBE-PARCB, PARCB, and patient-derived SCLC and SCPC lines. (**D**) Gene Ontology terms for hyperaccessible peaks in lung and prostate SCNCs. (**E**) TF motifs identified by HOMER motif analysis were plotted by ranks generated from their associated differential *P* values (table S7). (**F**) (Top) Selected visualization of ATAC-seq footprint. bp, base pair. (Bottom) The transcriptional activities were measured by gene signature scores (see materials and methods). Medians with interquartile ranges are shown. ***P* < 0.001, *****P* < 0.0001 (one-way ANOVA).

vival and initiation (*27–29*). NEUROD1, NEUROG2, and NKX homedomain TFs are implicated in neural stem cell patterning and neural progenitor fate specification (*30*). NKX2.1/TTF-1 is a biomarker for SCLC and SCPC (*31*). We found that the chromatin regions for binding p53 family TFs (p53, p63, and p73) and ETS family TFs (ELF3, ELF5, and ERG) were less accessible in SCNCs than in normal epithelial cells (Fig. 4E). Both of these TF families also play critical roles in neuronal development and lineage decisions (*32, 33*). Our motif analysis identified TFs known to be involved in neural proliferation and differentiation as well as additional TFs (table S7) with unexplored roles in SCNC. We next evaluated both RNA-seq and ATAC-seq data to compare the transcriptional activity of these TFs with the global accessibility of their DNA binding motifs. Our integrated analysis revealed that the differentially accessible TF motifs we identified correlated with the transcriptional output of TF target genes for each motif (Fig. 4F and fig. S14). These findings indicate that concomitant enrichment of proneural TFs and NKX homeodomain TFs and silencing of ETS family TF and p53 family TF transcriptional activities are conserved in and are likely critical for transformation to SCNC across tissue types.

Current therapeutic strategies designed to inhibit oncogenic pathways driving malignant phenotypes almost inevitably lead to treatment-resistant cancers. Increasingly, treatment-resistant tumors that assume aggressive clinical characteristics and molecular features of both stemlike and neuroendocrine lineages are being identified in a variety of epithelial cancers. We provide cellular, transcriptional, and epigenetic evidence that the SCNC phenotype that arises from distinct epithelial organs represents a common point in the evolution of cancers that is induced by shared genetic and epigenetic processes. Although normal human epithelial cells derived from developmentally distinct organs have their own molecular landscapes, our findings demonstrate that a defined set of oncogenic factors can induce the development of a common lethal neuroendocrine cancer lineage (SCNC) from different epithelial cell types. We have characterized the essential contribution of these factors and the convergence of both the transcriptional and chromatin landscapes during the process of transformation to SCNC. Our integrated molecular analyses have identified a group of key TFs that appear to be critical for the initiation and maintenance of SCNC, independently of the tissue of origin. These data may help inform efforts to identify novel therapeutic approaches for preventing the emergence of SCNCs and for treating them once they arise.

and SCLC lines exhibited similar chromatin accessibility landscapes (Fig. 4C and fig. S12), which paralleled their transcriptional resemblance. Notably, the chromatin accessibility profiles of patient-derived SCPC and SCLC cell lines clustered with our prostate and NHBE-PARCB cell lines but not with non-neuroendocrine prostate and lung epithelial cells (Fig. 4C). We identified shared hyper- and hypo-accessible chromatin regions in the lung and prostate SCNCs compared with normal lung and prostate epithelial cells (fig. S13). The SCNC-specific accessible regions were enriched for genes important for development and neuronal differentiation (Fig. 4D and table S6).

To gain insight into the possible biological roles of TFs that serve as master regulators in SCNCs, we performed DNA binding motif enrichment analysis of the hyperaccessible regions. This analysis revealed motifs corresponding to proneural TFs (ASCL1, NEUROD1, NEUROG2, and OLIG2) and NKX homeodomain TFs (NKX2.1, NKX2.2, NKX2.5, and NKX6.1) (Fig. 4E and table S7). ASCL1 and NEUROD1 are required for SCLC sur-

REFERENCES AND NOTES

1. D. Hanahan, R. A. Weinberg, *Cell* **144**, 646–674 (2011).
2. R. Nadal, M. Schweizer, O. N. Kryvenko, J. I. Epstein, M. A. Eisenberger, *Nat. Rev. Urol.* **11**, 213–219 (2014).
3. A. F. Gazdar, P. A. Bunn, J. D. Minna, *Nat. Rev. Cancer* **17**, 765 (2017).
4. L. V. Sequist *et al.*, *Sci. Transl. Med.* **3**, 75ra26 (2011).
5. H. L. Tan *et al.*, *Clin. Cancer Res.* **20**, 890–903 (2014).

6. M. J. Niederst et al., Nat. Commun. **6**, 6377 (2015).
7. H. Beltran et al., Nat. Med. **22**, 298–305 (2016).
8. S. Y. Ku et al., Science **355**, 78–83 (2017).
9. R. Fisher et al., Genome Biol. **15**, 433 (2014).
10. H. Chen, X. He, Mol. Biol. Evol. **33**, 4–12 (2016).
11. N. McGranahan, C. Swanton, Cell **168**, 613–628 (2017).
12. T. Stoyanova et al., Proc. Natl. Acad. Sci. U.S.A. **110**, 20111–20116 (2013).
13. V. Parimi, R. Goyal, K. Poropatich, X. J. Yang, Am. J. Clin. Exp. Urol. **2**, 273–285 (2014).
14. J. K. Lee et al., Cancer Cell **29**, 536–547 (2016).
15. Y. Lin, J. Fukuchi, R. A. Hiipakka, J. M. Kokontis, J. Xiang, Cell Res. **17**, 531–536 (2007).
16. M. Krajewska et al., Am. J. Pathol. **148**, 1567–1576 (1996).
17. L. D. Shultz, F. Ishikawa, D. L. Greiner, Nat. Rev. Immunol. **7**, 118–130 (2007).
18. P. Mu et al., Science **355**, 84–88 (2017).
19. M. Zou et al., Cancer Discov. **7**, 736–749 (2017).
20. E. Y. Chen et al., BMC Bioinformatics **14**, 128 (2013).
21. J. George et al., Nature **524**, 47–53 (2015).
22. National Cancer Institute, The Cancer Genome Atlas; https://cancergenome.nih.gov/.
23. H. Zhang et al., Mod. Pathol. **18**, 111–118 (2005).
24. J. Barretina et al., Nature **483**, 603–607 (2012).
25. W. K. Cheung, D. X. Nguyen, Oncogene **34**, 5771–5780 (2015).
26. W. A. Whyte et al., Cell **153**, 307–319 (2013).
27. H. Osada, Y. Tatematsu, Y. Yatabe, Y. Horio, T. Takahashi, Cancer Res. **65**, 10680–10685 (2005).
28. T. Jiang et al., Cancer Res. **69**, 845–854 (2009).
29. J. K. Osborne et al., Proc. Natl. Acad. Sci. U.S.A. **110**, 6524–6529 (2013).
30. F. Guillemot, Development **134**, 3771–3780 (2007).
31. S. N. Agoff et al., Mod. Pathol. **13**, 238–242 (2000).
32. P. Remy, M. Baltzinger, Oncogene **19**, 6417–6431 (2000).
33. B.Joseph, O. Hermanson, Exp. Cell Res. **316**, 1415–1421 (2010).
34. S. Heinz et al., Mol. Cell **38**, 576–589 (2010).

SUPPLEMENTARY MATERIALS

SCIENCE TRANSCENDING BOUNDARIES

ANNUAL MEETING
Washington, DC | Feb. 14–17, 2019

aaas.org/meetings

SHOWCASE YOUR RESEARCH WITH AN E-POSTER

E-poster abstract submissions are being accepted for the 2019 AAAS Annual Meeting General E-poster Session and Student E-poster Competition.
Deadline to submit is October 18, 2018.

E-POSTER CATEGORIES

Brain and Behavior | Cellular and Molecular Biology | Education

Developmental Biology, Physiology, and Immunology | Social Sciences

Medicine and Public Health | Physical Sciences | Science in Society

Environment and Ecology | Technology, Engineering, and Math

Meeting program and advance registration are available online.

Sample Preparation Cartridges and Plates

Waters Oasis PRiME MCX cartridges and 96-well plates selectively retain and concentrate basic compounds while removing up to 99% of phospholipids, using methods up to twice as fast as those for traditional mixed-mode, solid-phase extraction products. With the removal of phospholipids—the most abundant cause of matrix interference—from biological matrices, scientists will simplify their analyses, improve method robustness, and increase instrument uptime, while getting the accurate information they need to make informed decisions. Oasis PRiME MCX is a mixed-mode (reversed-phase and cation-exchange) sorbent that addresses the inherent complexities in quantifying target analytes in biological matrices, such as serum, plasma, whole blood, or human and animal tissue, and food samples including milk, meat, and eggs. No conditioning or equilibration steps are required prior to their use, saving scientists significant time and effort.

Waters
For info: 800-252-4752
www.waters.com/prime

Variable Power LED Reactor

The Uniqsis PhotoSyn is a new, high-power LED light unit that enables scalable flow photochemistry reactions, obtaining yields ranging from milligrams to hundreds of grams per day. It uses a pair of enclosed LED arrays to concentrate light inwards onto a central coil reactor. Each array is composed of 260 individual 1-W blue (455-nm) LEDs, making the unit particularly suitable for photoredox applications. These large-format arrays provide sufficient light to facilitate reactions in coil reactors (up to 60 mL), allowing work at a larger scale. Reactions may be run from 150°C down to subambient temperatures, depending upon the power output required. The variable power supply allows the power output to be adjusted from 10%–100%, making the unit suitable for both small-scale R&D and scale-up applications.

Uniqsis
For info: +44-(0)-845-864-7747
www.uniqsis.com

Light-Sheet Microscope for Cleared Tissue

The ct-dSPIM from Applied Scientific Instrumentation (ASI) is a flexible and easy-to-use implementation of selective plane illumination microscopy (SPIM), which allows for dual views (d) of large samples such as cleared tissue (ct). When used with ASI and Special Optics' new objective optimized for light-sheet imaging of cleared tissue, imaging depth can exceed 5 mm into flat samples. The objective accommodates media refractive indices (RI) from 1.33 to 1.56 (aqueous or organic media). The ct-dSPIM has successfully imaged various cleared-tissue samples from microtome-cut slices to whole mouse brains. As a dual-view system, the roles of the objectives can then be reversed to collect another stack from a different perspective; computationally merging the two stacks yields a 3D dataset with ~2x improvement in axial resolution.

Applied Scientific Instrumentation
For info: 800-706-2284
www.asiimaging.com

Electrofusion and Electroporation System

The ECM 2001+ is a multifunctional electrofusion and square-wave electroporation generator. The ability to generate both AC and DC waves allows for fast, efficient cell fusion in hybridoma production, hybrid-cell formation, and nuclear transfer applications. Electrofusion experiments with the ECM 2001+ are scalable from a few cells in 20-mL microslides all the way up to tens of millions of cells in 9-mL coaxial production chambers. For mammalian transfection applications, this system is powerful enough to yield high transfection efficiencies for difficult-to-transfect sample types, including stem cells, primary cells, mammalian tissues, and embryos. The ECM 2001+ is compatible with the full range of BTX specialty electrodes for in vitro, in vivo, ex vivo, in utero, and in ovo electroporation applications. Applications of the ECM 2001+ include cell fusion, hybridoma production, nuclear transfer, embryo manipulation, mammalian transfection, and CRISPR gene editing.

BTX
For info: 800-272-2775
www.btxonline.com

Micro-Flow Imaging System

Micro-Flow Imaging (MFI) combines the direct-imaging capabilities of digital microscopy with the precise control of microfluidics. What does that get you? High-resolution images with 85% sampling efficiency, more precise counts and sizing with full morphological detail for all subvisible particles in your sample, and the complete confidence that you can accurately identify every possible particle type—from protein aggregates to air bubbles. Images of the sample are captured as it passes through the flow cell's sensing zone. Every particle in every image is then analyzed to create a database of particle count, size, transparency, and morphology (or shape). And you'll have visual verification on the spot, as images are displayed in real time. You can also display results for many samples at once, making it easy to monitor stability and comparability. If you're looking for broader, deeper analysis of your biotherapeutic products, MFI is the particle-analysis tool for you.

ProteinSimple
For info: 888-607-9692
www.proteinsimple.com/mfi_5000.html

CRISPR RNA Libraries

Horizon Discovery's Edit-R CRISPRa arrayed crRNA (CRISPR RNA) libraries offer a powerful tool for drug discovery, pathway analysis, and disease progression studies. Unlike pooled lentiviral screens, Edit-R synthetic crRNA arrayed libraries enable one-gene-per-well investigation, using high-content assays to answer more in-depth biological questions. A CRISPRa-based system for overexpression studies also overcomes many of the shortcomings of early generations of DNA plasmid-encoded gene expression tools. By triggering the endogenous gene's expression, a gene is transcribed in its native form, so researchers can be assured of highly relevant results in their cell system. The Edit-R CRISPRa portfolio includes catalog libraries for popular human and mouse gene families, such as ubiquitin enzymes, transcription factors, and kinases, in addition to druggable gene targets and the whole human genome. Bespoke collections are also available to support researchers working with a specialized gene-target list.

Horizon Discovery
For info: 800-235-9880
www.horizondiscovery.com

Electronically submit your new product description or product literature information! Go to www.sciencemag.org/about/new-products-section for more information.

Newly offered instrumentation, apparatus, and laboratory materials of interest to researchers in all disciplines in academic, industrial, and governmental organizations are featured in this space. Emphasis is given to purpose, chief characteristics, and availability of products and materials. Endorsement by *Science* or AAAS of any products or materials mentioned is not implied. Additional information may be obtained from the manufacturer or supplier.

JOIN US.

FACULTY POSITIONS AT THE JACKSON LABORATORY

The Jackson Laboratory is inviting applications for Assistant, Associate, and Full Professors in Bioinformatics and Computational Biology research at The Jackson Laboratory for Mammalian Genetics in Bar Harbor, Maine.

The Jackson Laboratory (JAX) is an independent, nonprofit biomedical research institution and features a National Cancer Institute-designated Cancer Center. JAX is a highly collaborative environment that fosters multidisciplinary strategies to investigate complex biological questions. Our faculty use genetic and genomic approaches to conduct basic and translational research on fundamental biological processes in disease using mouse models and human studies. JAX is home to internationally recognized information resources including the Mouse Genome Informatics (MGI) database, GeneWeaver integrative genomics tools, and the Gene Ontology and Human Phenotype Ontology projects.

Successful candidates must hold a Ph.D. and/or M.D. degree; have two to five years of postdoctoral training; have a record of significant research accomplishments; and present long-term research goals consistent with establishing and/or maintaining a highly successful research program.

Applicants should submit a curriculum vitae and a concise statement of research interests and plans, through the JAX Careers portal at *www.jax.org/careers*. **In addition, please arrange to have three letters of reference sent to:** *facultyjobs@jax.org*.

The Jackson Laboratory provides equal employment opportunities to all employees and applicants for employment in all job classifications without regard to race, color, religion, age, mental disability, physical disability, medical condition, gender, sexual orientation, genetic information, ancestry, marital status, national origin, veteran status and other classifications protected by applicable state and local non-discrimination laws.

PRIMARY RESEARCH AREAS OF RECRUITMENT

— Systems genetics
Systems approaches to complex trait analysis, including genetic analysis of transcript expression, histone modifications and protein abundance, as well as deep learning and neural network analysis of high-volume genomic, imaging, video and other whole organism data.

— Cross-organism bioinformatics
Initiatives include creating and aligning animal models of human disease with clinical genetics by using algorithms to translate results for diverse genomes and phenomes that maximize translational potential of animal models.

— Computational biology of cancer
Focus areas include models of cancer biology, tumor evolution, tumor-immune systems interactions and mechanisms of therapy response and resistance.

The Jackson Laboratory
Leading the search for tomorrow's cures

CHALLENGE
KNOWLEDGE

NTNU is Norway's leading science and technology university. Here, equality, stability and long-term government research funding create freedom to innovate, collaborate and explore new ideas. Together, we challenge — challenge established truths, naysayers and ourselves. We create knowledge for a better world.

www.ntnu.edu

NTNU

Knowledge for a better world

Cluster Hiring of Clinical Medicine + X

As a top-level comprehensive research university in China, Peking University (PKU) assumes the central mission of cultivating talents that will lead the future, and generate new thoughts, frontier science and future technology that promote the progress of the state and mankind. It always orients its scientific research in a way that takes into account the international frontiers of science and the state's major strategic demands, and has established itself, with preliminary success as a high-level research center that integrates national fundamental research, high-tech research and medical and health technology research. PKU is home to a large group of talents, thanks to its effort in cultivating and recruiting many first-class scientists, discipline leaders and innovative teams that stay active at international academic frontiers and satisfy major national strategic demands. With its great academic environment, spirit of free exploration and pursuit of excellence, the campus is attracting more and more talents and is becoming known as a place boasting of a concentration of great masters, generations of renowned scholars as well as high morals.

With the coming of its 120th anniversary, PKU constructed in particular a "cluster" of several teams for interdisciplinary research, and take the lead to kick off the Cluster Hiring of Clinical Medicine + X. The successful establishment of the cluster relies on the wide range of disciplines available at PKU. Using clinical needs as its guidance, the cluster will break down disciplinary barriers in a way that emphasizes the fusion, collaboration, and complementation between different disciples. It will solve major clinical problems of general interest in such fields as innovative drug, medical technology and biomedical engineering, multi-omics of precision medicine, healthcare Big Data and intelligent healthcare. It will also improve clinical diagnosis and treatment levels. What's more, it will cultivate innovative compound interdisciplinary talents through cross and joint innovative studies between clinical medicine and basic medical science, pharmaceutics, bioscience, engineering sciences, information science and material science, with the emphasis on interdisciplinary integration, collaborative innovation and complementation of advantages.

Two kinds of people are particularly welcome to apply: scholars with a strong research and teaching record who are engaged in basic studies closely related to clinical medicine and proficient in common key technology and cutting-edge core technology of clinical science; have made dramatic breakthroughs in relevant fields, and are expected to develop innovative theories; physicians who are certified to practice medicine in Mainland China who are very interested in and devoted to basic studies, have made dramatic breakthroughs in relevant fields and established new techniques, theories or doctrines, and are proficient in core theories, projects or central techniques that can lead the development of relevant medical sectors. We're sincerely inviting famous scholars, academic leaders, and promising young scholars in various interdisciplinary areas to join us.

The applicant should have obtained a doctor's degree in relevant orientation, has had a solid research foundation and outstanding research findings in said fields and great potential for being an excellent teacher and tutor, and is also an excellent team player in interdisciplinary collaboration.

PKU will place the successful applicant in a proper position, at PKU, PUHSC or an affiliated hospital according to his/her academic background, research interest and teaching plan, and provide him/her with competitive post salary, sufficient start-up fund, necessary laboratory and office space and other support. During the employment, PKU will evaluate the professional development capacity and potential of those teachers and provide them with the best opportunities and approaches of vocational development in line with international conventions and common practices in the academic circles.

All positions will remain open until filled. Application materials should include a resume, 3 letters of recommendations from peer experts, the statement of research and teaching plan and other proof materials. For more details, please refer to the recruiting website (http://xkb.pku.edu.cn/).

Contact of PKU:
Xiaopeng Liu,
Tel: 86-10-62755591,
Email: lxp@pku.edu.cn

Contact of PUHSC:
Yuan Cao,
Tel: 86-10-82805626,
Email: caoyuan9004@bjmu.edu.cn

北京大学人民医院
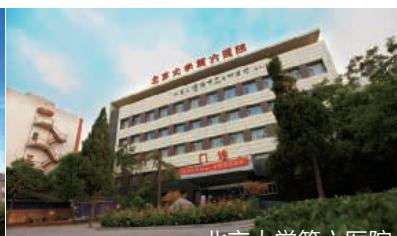
北京大学第六医院

北京大学第一医院

北京大学肿瘤医院

北京大学第三医院

北京大学口腔医院

摄影：黄政翔

High-level Talents Recruitment
Xi'an University of Science and Technology

Time and place:
► ESSCA Budapest, Budapest, Hungary,
   15:00-18:00 Oct. 18, 2018
► Jagiellonian University, Krakow, Poland
   9:00-16:00 Oct. 20, 2018
► Boerno Exhibition Center, Brno, Czech Republic
   9:00-16:00, Oct. 23, 2018

Xi'an University of Science and Technology (XUST) is a university co-sponsored by the State Bureau of Production Safety Supervision and Management and the People's Government of Shaanxi Province. As one of the key high-level universities in Shaanxi as well as in the middle west China, XUST is an important training base and scientific research center which cultivates professionals of geology，mining，safety science and engineering. Located in Xi'an, the capital city of Shaanxi Province, XUST has two campuses: the head one located in Yanta District and the new one in Lintong District. XUST owns 18 colleges and departments, 6 post-doctoral centers, 7first-level disciplines and 40 second-level disciplines to grant doctoral degree, 25 first-level disciplines and 107 second-level disciplines to grant master's degree, 18 cultivation programs on Master of Engineering, 1 MBA program on master's degree and 1 MPAcc program on master's degree as well as 56 undergraduate programs. Featured in Mining, Geology, and the relevant disciplines, it has now grown into a well-rounded university with 23,000 full-time students, integrating the primary subject of engineering with subjects of science, literature, management, law, economics and art.

Fields of recruitment:
● Energy and mining
● Mechanics
● Materials, Physics, Chemistry & Mathematics
● Electronics, Information & Control
● Environment, Architecture and Civil engineering
● Management, Accounting and Economics
● Humanities & Social Science

Requirements of recruitment:
1. Applicants under the age of 45 who are candidates of the National Recruitment Program for Young Professionals, National Special Support Program for Outstanding Young Scholars, Yangtze River Scholar Program for Young Scholars, New Century Talent Supporting Project, Hundred Talents Program; Project leaders of State Natural (Social) Sciences Foundation.

2. Applicants under the age of 40 who are leaders of provincial and ministerial key disciplines, winners of science and technology achievement as the head of the project; young scientific leaders of provincial level or of the equivalent level; candidates of provincial recruitment programs; leading talents of any discipline, members of CAE or CAS teams.

3.Distinguished doctors under the age of 32 with at least one year's experience of studying abroad whose undergraduate, postgraduate and doctoral study are accomplished in domestic and overseas high-level universities or national scientific research institutions. Research experience in national research projects and publication of 5 articles in core journals (3 in SCI or 1 in SSCI / CSSCI) are also expected.

We provide competitive packages of salary and benefits for talents of different levels including free apartments, settling-in allowance, research/lab fund, etc.

Contact us:
E-mail: xkdrcb@xust.edu.cn
(Please name your CV attachment in the following format: name-specialized field-school.)
WeChat: 18991876761 / 78024224
Mobile: +86 18991876761 / +86 13720555324

For more information, please refer to our website https://www.xust.edu.cn/.

The Huge Talent Dividend and Demand Spurred by the Belt and Road Initiative

The remarkable achievement of opening an overland route linking the East and the West by Zhang Qian more than 2,000 years ago—known as the "Silk Road"—has had far-reaching consequences. The legend of Xuan Zang, who traveled the Silk Road to India in the seventh century, still resonates with the Chinese people. And the impressive feat of Zheng He, who opened the "Maritime Silk Road" with his seven voyages to the Western oceans, still inspires after more than 600 years. From the Flying Devi (angel) murals of the Mogao Grottoes in Dunhuang, which represent the integration of Chinese, Indian, and ancient Western culture, to the discovery of the Belitung shipwreck—an Arabian ship fully loaded with Chinese gold and ceramics—found in the Java Sea, the ancient Silk Road (on both land and sea) witnessed the courage of our intrepid Chinese ancestors and recorded the beautiful history of integration and common development between East and West.

As time went by, we saw the once-desert land of the Silk Road connected by railways, and human-powered ships that sailed its ocean routes replaced by steam and gas-powered vessels—and now the Silk Road itself is beginning to revive. In the autumn of 2013, Chinese president Xi Jinping travelled west to Kazakhstan and south to Indonesia and proposed the major initiatives of the Silk Road Economic Belt and the 21st-Century Maritime Silk Road. From then on, China's commercial center started to move back to where it originated 2,000 years ago—the Silk Road.

This year marks the fifth anniversary of the Belt and Road Initiative. Over the past five years, China has set up 82 overseas economic cooperation zones along Belt and Road corridors, with a total investment of USD 28.9 billion, having created 244,000 jobs. As of May 2018, China has signed 16 free-trade agreements with 24 countries across four continents, nearly half of which are along the "Belt and Road."

Paving the way for a world-class educational

Education and talent have played a fundamental and pioneering role in the work of building the Belt and Road. China's efforts to strengthen communication, cooperation, and education with countries along the Belt and Road not only play an important part in the initiative, but also give a strong boost to its talent programs. China's proposal to build a Belt and Road educational community, and the claim that

Shixin Wang
Deputy Chief Editor of China Education Online, Chief Executive Editor of AcaBridge

the Belt and Road is the roadmap for China to become a world-class educational center, have received enthusiastic responses from Chinese universities and colleges, which hope to seize this opportunity for progress, and are also an important part of the initiative.

Since the Belt and Road Initiative began, Beijing Foreign Studies University has created more than 30 programs to teach languages spoken by countries along the Belt and Road, thus fulfilling domestic needs for relevant language skills.

Xinjiang and Fujian: Silk Road trailblazers

Xinjiang and Fujian are the respective cores of the Silk Road Economic Belt and the 21st-Century Maritime Silk Road. The once-backward Xinjiang is now at the forefront of the initiative: Xinjiang University and other universities in the area have now established 10 branches of the Confucius Institute in countries such as Russia, Kyrgyzstan, Tajikistan, Kazakhstan, Pakistan, and Mongolia, with 36,000-plus students enrolled, and nearly 100,000 people engaged in learning the Chinese language and experiencing Chinese culture.

Furthermore, Fujian is trying to promote international communication and cooperation between universities in countries along the Belt and Road through the University Alliance of the New Maritime Silk Road. Xiamen University opened a branch in Malaysia in February 2016, which has enrolled over 3,000 students from 17 countries. Huaqiao University's Chinese-language teaching program for officials from countries along the Belt and Road has trained hundreds of students. Fujian Normal University has also reached an agreement with Universitas Al Azhar Indonesia and Angeles University Foundation to jointly apply for various Belt and Road cooperation and exchange programs for building culture and improving education. More close exchanges with universities from East and South Asia are expected.

Creating educational alliances

The University Alliance of the Belt and Road, first launched by Lanzhou University, has seen its members increase from 47 institutions in eight countries to 126 institutions in 25 countries spread across Asia, Europe, Africa, and America. Gansu University of Chinese Medicine has established eight branches of the Qihuang Academy of Traditional Chinese Medicine and four centers of Chinese medicine, in collaboration with universities from many countries, including Russia, Ukraine, Kyrgyzstan, France, Moldova, Hungary, Madagascar, and New Zealand.

Established by Shaanxi Normal University and with the support of dozens of universities and scientific research institutes from countries and districts along the Belt and Road, three alliances—the Silk Road Teacher Education Alliance, the Silk Road Humanities and Social Sciences Alliance, and the Silk Road Library, Archives and Publishing Alliance—aim to set up a Silk Road platform for exchange and cooperation and promote common development among member states in humanities education, along with collaborative efforts to open schools, organize student exchange programs, and advocate a credit system shared by all collectively run universities. Following the Belt and Road strategy of development, Hefei University of Technology now offers a degree course in the Indonesian language, and has turned to AcaBridge for help in attracting talent with language skills. AcaBridge, in turn, found the best candidates from hundreds of thousands of experts in its database to fulfill the university's need.

Establishing industry-based and cooperatively run universities

Industry-based universities are another crucial part of the economic-development component of the Belt and Road Initiative. Dozens of industry-based universities, such as Ningbo Polytechnic, various businesses, and industry associations from Chinese provinces and cities along the Belt and Road have formed an alliance of educational and industrial organizations to further promote cooperation in vocational education and explore new models for international vocational universities.

Programs for Sino–foreign, cooperatively run schools have grown rapidly. As of May 2017, a total of 2,539 cooperatively run schools have been authorized: 1,248 of these are able to confer Bachelor's degrees and above, and 928 are industry-based universities. Fifteen cooperatively run institutions, such as Shenzhen MSU-BIT University and the Zhejiang University–University of Edinburgh Institute (ZJE), have also been established. Meanwhile, programs for opening schools abroad are also steadily moving forward. As of 2016, four institutions and 98 school programs have been created overseas—and most of them are spread along the Belt and Road.

Searching for new talent

The rapid development of the Belt and Road Initiative is bound to bring about both a huge talent dividend and a demand for new talent. Thus, the continuous growth of this project entails much investment in training talent with global perspectives, business knowledge, cultural awareness, and language skills, to ensure that China can take full advantage of the initiative and utilize global human resources to foster international communication and cooperation on an extensive scale.

We welcome excellent scholars at home and abroad who are interested in applying for programs related to the "Belt and Road" Initiative to contact us through AcaBridge (consultant@acabridge.edu.cn), which provides one-on-one consultations. Further information can be found at www.edu.cn/jjydyl.

XINYU UNIVERSITY

Located in Xinyu City, Jiangxi Province, China, Xinyu University is a public full-time undergraduate university occupying an area of 33 acres (or 2000 mu). On the 387,000 square meters building area, Xinyu University has built 162 campus experiments rooms, 163 stable external training bases, together with a large school library with 1.527 million volumes of books and 27 Chinese and foreign language databases in it. Xinyu University has had 13 secondary schools, 69 majors and specialists, covering seven categories including engineering, literature, science, medicine, education, management, and arts. Adhering to the school motto of "Seeking Creative, Practical, and Perfect", Xinyu University vigorously promotes the socialist core values. Through actively establishing a series of campus culture of "Integrity, Inclusiveness, Nobility, and Harmony" and creating "Classic Nurturing", "Square Culture at

Weekends", "Living Library", and "One School, One Feature", Xinyu University has achieved remarkable results in campus culture construction, for which the university was honored as the first ten civilized campuses of colleges and universities in Jiangxi Province in 2017.

Demand Specialty

Mechanical Engineering, Management Science and Engineering, Mechanics, Civil Engineering, Architecture, Computer Science and Technology, Mathematics, Philosophy, English language and literature, News communication, Drama and film and television, Applied economics, Business management, Statistics, Clinical medicine, Basic medicine, Pharmacology, Arts.

Remuneration and living treatment

Doctors from overseas universities:
1. House purchase subsidies: 400 thousand yuan; or provide a set of

housing (blank room 120, after 6 years of work in school, property rights to their own all) and 200 thousand Yuan decoration subsidies; signing a contract the subsides will be given 50%, the remaining by the year;
2. A one-time home settling fee: 150 thousand;
3. Funding for scientific research start-up: 50 thousand yuan for liberal arts, 100 thousand yuan in Medical Sciences and 150 thousand yuan in engineering, 25% for scientific research related bills, and 15% by 15% votes per year.
4. Doctoral degree allowance: 1500 yuan / month during the service period. After the expiration of the service, refer to the stipulation of the degree stipulation introduced by the school at that time.
5. Doctors not obtained the title of the secondary high school:
After signing an agreement with the school, the candidates will be appointed as associate professor,

and enjoys the basic and rewarding performance wages in the school according to the title of associate professor.
6. All doctors will be equipped with a doctoral workshop and a computer, spouses can be adjusted and arranged for work within the school.
7. Doctors recruited also enjoy the treatment provided by Xinyu city: If doctors have worked for three years, the municipal government award a subsidy of 250 thousand yuan and living allowance of 2000 yuan / month (specific to the city documents), and can rent the city's high level talent apartment for free for three years (live only carrying some necessities); if children are in the compulsory education stage, you own the right of school selection.

Contact: Liu Pingping
Contact phone: 0790-6666026
Mailbox: xyxyrsk@163.com

Traditional Chinese Medicine for people's health along the "Belt and Road"

Traditional medicine, like modern medicine, is an important method to maintain human health, especially in developing countries along the "Belt and Road". According to the report of the World Health Organization (WHO), about 80% of the world's population has been used or using traditional medicine to prevent and treat disease. Traditional medicine is considered as an essential pillar to achieve the goal of "health for all" of WHO. Traditional Chinese Medicine

TCM is one of the most widely used traditional medicine, which was formed and developed in the daily

life of the people and in the process of their fight against diseases over thousands of years. In China, there is about 3960 hospitals of TCM, which can provide clinical service for 910 million visits each year. The combination of TCM with modern science and technology can produce original achievement. Youyou Tu led a team that transformed an ancient Chinese healing method into the most powerful antimalarial medicine--artemisinin, which has saved hundred million people in developing countries and won the 2015 Nobel Prize in Physiology or Medicine.
Nowadays, TCM has been spread to 183 countries or regions around the world. The Chinese government has signed 86 TCM cooperation agreements with other countries and international organizations, and has supported the building of 10 TCM clinical centers overseas. Chinese medical products have gradually entered the international medical market and some of them have been registered in Russia, Cuba, Vietnam, Singapore, United Arab Emirates, and other nations. In addition, about 30 countries or regions have opened a couple of hundred TCM schools to train native TCM practitioners; each year, there are about 15000 oversea students learning TCM in China. There are about 200,000 foreign

patients accepted TCM treatment in China. TCM has becoming a brilliant name card to strengthen People-to-people Ties in countries or regions along the "Belt and Road". Tianjin University of Traditional Chinese Medicine (TUTCM), established in 1958, is one of the leading university in TCM clinical service, scientific research and education, especially international collaboration. TUTCM is one of the "Educational Aid Center for Africa and Asia" and "China-ASEAN Educational Training Center" of Ministry of Education (MOE), TCM International Cooperation Center of State Administration of Traditional Chinese Medicine (SATCM), and Belt & Road Training Center for International Education of TCM. TUTCM is one of the chief commissioner of teaching committee of MOE and World Federation of Chinese Medicine Societies (WFCMS). In 2017, TUTCM was selected into China "Double First-class" university with "Double First-class" discipline.
There are over 2,000 international students from more than 60 countries or regions. Collaborations and academic exchanges have linked over 146 universities, research institutes and medical centers in 41 countries or regions.
On September 16 this year, the

5th World Education Conference of Chinese Medicine and the 60th anniversary were successfully held in the new campus of TUTCM, which covers an area about 1.7km2. There were more than 500 representatives from more than 32 countries or regions attended the conference. In order to provide more and higher quality TCM service for people from countries or regions along the "Belt and Road", TUTCM strengthens its international education capability by establishing the WFCMS "Belt and Road" training based for foreign teachers of TCM and distance education training center in September 2018. We believe, the fast development of TCM not only can bring benefit to Chinese people, but also for worldwide people's health.

Junhua ZHANG, Lirong CHU
Tianjin University of Traditional Chinese Medicine
Email: zjhtcm@foxmail.com

DATA SCIENCE
GENOME BIO-
LOGY HUMAN-
ITIES & ARTS
INFECTION
BIOLOGY
NANO SCIENCE
& MOLECULAR
ENGINEERING
SOCIAL SCIENCES
SUSTAIN
ABILITY

JOIN US

Unlikely integrations of deep expertise can pave the way to new discovery. That is why Cornell University is developing a research community that combines strategic discipline areas that point the way toward the solutions of tomorrow. Join our team of world-class experts who are changing the landscape of research innovation.
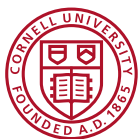
Cornell University

cornell.edu/collaborate

Faculty Positions in Cancer Cell and Molecular Biology
Purdue University

40 YEARS OF CANCER RESEARCH

BIO PURDUE

The Department of Biological Sciences in the College of Science at Purdue University, in conjunction with the NCI designated Purdue Center for Cancer Research, invites applicants for one or more tenure-track faculty positions in Cancer Cell and Molecular Biology. Research programs with emphasis in the following areas will be considered:

Cancer Cell Biology: • Cancer stem cell biology including model systems employing iPS cell technology • Signaling pathways impacting tumor development • Tumor microenvironment • Tumor immunology and immunotherapy

Cancer Molecular Biology: • Control of gene expression in cancer cells by transcription factors, epigenetics, or 3D/4D organization of the genome • Post-transcriptional control of cancer-related processes by non-coding RNAs • Tissue reprogramming and differentiation of cancer stem cells

Growth in Cancer Biology is part of a large-scale interdisciplinary hiring effort across key strategic areas in the College of Science and comes at a time when the College is under new leadership with multiple commitments of significant investment. Opportunities for collaboration exist across the Purdue campus and in the Purdue Center for Cancer Research (https://www.cancerresearch.purdue.edu/) as well as with the new campus-wide Data Science initiative in bio data sciences. Applicants must have a Ph.D. or equivalent in biology or related field and at least 2 years of postdoctoral experience. We expect to fill positions at the tenure-track, Assistant Professor level. Successful applicants are expected to conduct research in cancer cell biology or cancer molecular biology, teach undergraduate and graduate students, and participate in ongoing programs within the Department of Biological Sciences and the College of Science.

The Department of Biological Sciences has over 50 faculty members conducting research in cancer biology, structural biology, neurobiology, virology, microbiology, molecular and cell biology, bioinformatics, evolutionary biology and ecology, and biology education (https://www.bio.purdue.edu/). Research in the Department is central to a highly interactive community of scientists from the Colleges of Science, Agriculture, Pharmacy, Veterinary Medicine and Engineering and from interdisciplinary centers in the Life Sciences including the Purdue Institute for Inflammation, Immunology and Infectious Diseases (http://www.purdue.edu/discoverypark/pillars/pi4d/index.php), the Purdue Institute for Integrative Neuroscience (http://www.purdue.edu/discoverypark/pillars/integrative-neuroscience-center/index.php), and the Purdue Institute for Drug Discovery (http://www.purdue.edu/discoverypark/drug-discovery/).

Applications should be submitted electronically to https://hiring.science.purdue.edu/ as a single PDF file containing a cover letter indicating the position of interest, a detailed curriculum vitae, contact information for three references, a two-page summary of research interests, and a one-page teaching statement. Purdue University's Department of Biological Sciences is committed to advancing diversity in all areas of faculty effort, including scholarship, instruction, and engagement. Candidates should address at least one of these areas in their cover letter, indicating past experiences, current interests or activities, and/or future goals to promote a climate that values diversity and inclusion. Purdue University is an ADVANCE institution (https://www.purdue.edu/advance-purdue/). Direct inquiries to Elizabeth J. Taparowsky, Chair, Cancer Biology Search Committee at Search@bio.purdue.edu or **Cancer Biology Search Committee, Department of Biological Sciences, Purdue University, 915 W. State St., West Lafayette, IN 47907-2054**. Review of applications will begin October 8, 2018 and continue until the positions are filled. A background check is required for employment.

Purdue University is an EOE/Affirmative Action Employer. All qualified applicants, including minorities, women, individuals with disabilities and veterans are encouraged to apply.

THE UNIVERSITY OF NEW MEXICO.

The Department of Biology at the University of New Mexico is seeking two Assistant Professors, one in the area of Emerging Animal Models and one in Genomics/Bioinformatics.

Genomics/Bioinformatics: We are interested in applicants who use genomics and computational approaches to investigate fundamental questions in biology. We seek a colleague who will establish and maintain a vigorous, externally funded research program, which complements existing research in a broadly-based biology department. Candidates with expertise in computer science and statistics are encouraged to apply. UNM is strategically located near multiple national laboratories and has extensive resources for genomics research including molecular and cell biology core facilities and the Center for Advanced Research Computing. For complete job requirements, see Assistant Professor, Genomics/Bioinformatics at https://unmjobs.unm.edu, REQ#6226.

Emerging Animal Models: We are interested in a colleague focusing on species beyond the traditional model organisms to investigate fundamental questions in biology. These could include but are not limited to investigations of genome to phenome relationships, physiological adaptation, development, or host-microbe interactions at the organism or sub-organism level. We seek a colleague who can establish and maintain a vigorous, interdisciplinary, and externally funded research program that complements existing research in a broadly based biology department. For complete job requirements, see Assistant Professor, Emerging Models at https://unmjobs.unm.edu, REQ#6227.

For each position, the ideal candidate will be dedicated to excellence in teaching at the undergraduate through graduate levels at a minority-majority institution, including a demonstrated commitment to diversity, equity, inclusion, student success, as well as working with a diverse community. Applicants must have completed a Ph.D. in Biology or a related discipline at the time of application. Preference given to candidates with post-doctoral research experience at the start of the appointment.

The University of New Mexico carries the Carnegie Classification "Highest Research Activity". UNM is located in Albuquerque, a city with a rich cultural heritage situated in the spectacular Southwest.

Applicants should submit a cover letter, curriculum vitae, three representative reprints, description of current and future research and teaching interests, a diversity statement, and names and contact information of three references. All application materials must be uploaded and submitted through UNMjobs, https://unmjobs.unm.edu. For best consideration, application materials must be received by November 15, 2018.

The University of New Mexico is an Equal Opportunity/Affirmative Action Employer and Educator. Women and underrepresented minorities are strongly encouraged to apply.

International Research Leader Grants

novo nordisk fonden

EXCEPTIONAL GRANTS FOR EXCEPTIONAL SCIENTISTS

Two calls for applications within the areas of biomedicine & biotechnology

International Research Leader Grants from the Novo Nordisk Foundation are for outstanding scientists to establish and run their laboratories in Denmark.

Novo Nordisk Foundation
Laureate Research Grants

» Individual grants up to
**DKK 50 million over 7 years
(EUR ~6.7 million, USD ~7.9 million)**

» For principal investigators who have directed an independent research group for 7 or more years.

» Grant holders can apply for continued Laureate Research Grant funding, up to DKK 35 million over 7 additional years.

Novo Nordisk Foundation
Young Investigator Awards

» Individual grants up to
**DKK 25 million over 7 years
(EUR ~3.4 million, USD ~3.9 million)**

» For principal investigators who have directed an independent research group for less than 7 years.

» Award holders can apply for further funding from other Novo Nordisk Foundation grant programs.

Stage 1 application period for both calls

October 12, 2018 – November 15, 2018

To learn more, please visit: **novonordiskfonden.dk/en**

Massachusetts Institute of Technology

Department of Earth, Atmospheric, and Planetary Sciences

Faculty Position in Geophysics and Geochemistry

The Department of Earth, Atmospheric, and Planetary Sciences at the Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, invites qualified candidates to apply for a tenure-track faculty position. The search is in the broad area of geophysics and geochemistry encompassing the Earth and other planetary bodies in the solar system. We seek candidates who use theory, observation, and/or experimentation and particularly encourage applicants whose work crosses traditional disciplinary boundaries. Candidates should have the potential for innovation and leadership in research and a commitment to teaching at the undergraduate and graduate levels.

Applicants must hold a Ph.D. in geoscience or related field by the start of employment. Our intent is to hire at the assistant professor level, but more senior appointments may also be considered. A complete application must include a cover letter, curriculum vitae, one- to two-page descriptions each of research and teaching plans, and three letters of recommendation. We request that in their cover letter, applicants explicitly commit to our department's code of conduct: https://eapsweb.mit.edu/about/code-conduct

Applications are being accepted at Academic Jobs Online: https://academicjobsonline.org/ajo/jobs/11380

To receive full consideration, complete applications must be received by November 1, 2018.

Search Contact: Ms. Karen Fosher, HR Administrator, EAPS, 54-924, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139-4307, email: kfosher@mit.edu

MIT is an equal employment opportunity employer. All qualified applicants will receive consideration for employment and will not be discriminated against on the basis of race, color, sex, sexual orientation, gender identity, religion, disability, age, genetic information, veteran status, ancestry, or national or ethnic origin.

http://web.mit.edu

AUGUSTA UNIVERSITY

GEORGIA CANCER CENTER

Basic and Translational Cancer Research Scientists

Georgia Cancer Center at Augusta University in Augusta, GA is undergoing an unprecedented expansion in its basic, translational and cancer prevention programs as part of an initiative to achieve designation as an NCI Cancer Center. Following the successful completion of Phase 1 of this expansion, we invite applications for positions at **Assistant, Associate, and Full Professor** levels from interested individuals having expertise in **tumor immunology, tumor angiogenesis, cancer genetics and genomics, chaperone biology, cancer metabolism and experimental therapeutics**.

Applicants must have active extramural research funding (preferably from NCI) and a strong track record of independent research. Ideal candidates will have conducted research in an NCI-designated Cancer Center. Successful applicants will join one of our three collaborative programs that work closely with clinical research oncologists to promote translational research.

A competitive salary and start-up package will be provided, commensurate with experience and academic qualifications. A summary of research interests, curriculum vitae, and names of three references should be sent to **Khensley@augusta.edu**.

HARVARD MEDICAL SCHOOL

ASSISTANT PROFESSOR
Department of Immunology

The new Department of Immunology at Harvard Medical School invites applications for a tenure-track position at the rank of Assistant Professor. We seek an outstanding scientist performing basic and/or translational research in any area of innate or adaptive immunity. This position offers outstanding scholarly and scientific resources in a collegial and collaborative department with strong ties to related departments throughout Harvard University, the Harvard-affiliated teaching hospitals and the Boston immunology community. The position offers the opportunity to teach exceptional graduate and medical students with strong interests in immunology. The research space will be located in a modern and central research building at Harvard Medical School. Candidates must have a Ph.D., an M.D. or an equivalent graduate degree.

For consideration please upload to the website below a cover letter, a C.V. (including a list of publications), and a concise summary of research accomplishments and future interests. Three letters of recommendation will be required – to be submitted by the application deadline.

The application deadline is December 7, 2018.

Submit all materials via:
http://academicpositions.harvard.edu/postings/8452

We are an Equal Opportunity Employer and all qualified applicants will receive consideration for employment without regard to race, color, religion, sex, national origin, disability status, protected veteran status, gender identity, sexual orientation, pregnancy and pregnancy-related conditions or any other characteristic protected by law.

UNC SCHOOL OF MEDICINE
Computational Medicine Program

Tenure Track Faculty Positions

The newly formed Computational Medicine Program at the University of North Carolina at Chapel Hill (UNC-CH) invites applicants for two tenure-track faculty positions in Computational Medicine. The Computational Medicine Program at UNC-CH is a newly established research initiative that seeks to integrate modern computational approaches, including mathematical modeling, machine learning, and network analysis, with cutting-edge experimental techniques to address problems relevant to human health. While the area of study is flexible, an ideal candidate would apply methods from systems biology to investigate the molecular mechanisms dysregulated in human diseases, and to eventually develop predictive models of these disease states. Each recruit will hold a position, and research laboratory space, within the Computational Medicine Program, and also hold a position within a Department; these two searches include joint recruitments with the Department of Pharmacology, and a separate search with the Department of Genetics.

The successful candidates will have a doctoral degree and 3-6 years of postdoctoral experience in applying computational and quantitative approaches to problems in biomedical research. They would be expected to direct an independent research program supported by extramural funding, and participate in team science projects initiated through the Computational Medicine Program. They would also participate in graduate student mentoring and teaching at the graduate level. It is anticipated that some candidates may be entirely computational, while others will incorporate an experimental component into their research; both types are eligible and encouraged to apply. Assistant professors and associate professors are also welcome to apply.

Applicants should submit a cover letter, research statement, curriculum vitae and names of at least three references to: **http://unc.peopleadmin.com/postings/145663**.

Review of applicants will begin immediately and continue until the positions are filled.

RUTGERS

Environmental and Occupational Health Sciences Institute | EOHSI

The Environmental and Occupational Health Sciences Institute (EOHSI) at Rutgers, the State University of New Jersey is recognized as one of the premier institutes in the field of environmental and occupational health for over 30 years. EOHSI is home to multiple research centers including a P30 Center of Excellence, two T32 training grants (Toxicology and Exposure Science) from the NIEHS, a CounterAct Center, among others. EOHSI also houses an Occupational and Environmental Health Clinic, home to the Rutgers Occupational Medicine Residency Program and to a CDC Center for treatment and research on health issues in First Responders to the 9/11 World Trade Center attacks. EOHSI is a resource for occupational and environmental health at the local state and national levels. The Institute is undergoing a period of expansion and seeks outstanding scientists with experience in a broad range of disciplines related to environmental health including toxicology, exposure science and engineering, epidemiology, and occupational health to participate in research and teaching activities. For further information about EOHSI please visit our website http://eohsi.rutgers.edu/

Tenure Track Positions: We seek PhD, DVM, MD or equivalent degree applicants for tenure or tenure-track faculty positions at the assistant, associate and full professor levels with a record of RO1 (or equivalent) research grant funding. Faculty have access to state-of-the-art resources for genomic, epigenetic and proteomic analysis, next generation DNA sequencing, chemical analysis, exposure assessment, modeling, measurement of biological response indicators, trainee resources and pilot funding provided by our Centers and training grants.

Director of Exposure Science and Epidemiology Division: EOHSI is also seeking candidates with a doctorate in exposure science, epidemiology, or related disciplines and demonstrated history of Federal grant funding to lead our Exposure Science and Epidemiology Division. The Division supports exposure, exposome and epidemiologic research of exposed and vulnerable communities, including citizen science and human studies in our Controlled Exposure Facility. Sensor development for environmental monitoring, modeling of exposure using toxicology and human health effects data, and biomarker discovery linking exposure to health effects are examples of ongoing research within the Division

Applicants should send a letter of interest, C.V., statement of research and teaching interests, and names of three referees, in electronic format to:

Nancy Fiedler, Ph.D
EOHSI
170 Frelinghuysen Rd.,
Piscataway NJ, 08854.
nfiedler@eohsi.rutgers.edu

Rutgers University is an Equal Opportunity, Affirmative Action Employer. The University actively encourages applications and nominations of women, minorities, and persons with disabilities and applications from candidates with diverse cultural backgrounds.

Faculty Positions in Integrative Biosciences

Wayne State University (WSU) is recruiting up to15 faculty (open rank) for research and development programs as part of the continued expansion of a broad institutional initiative in Integrative Biosciences. This initiative leverages a new 200,000 sq. ft. Integrative Biosciences Center (IBio) that houses coordinated inter- and trans-disciplinary research teams, and a Clinical Research Center. Programmatic themes include a focus on pathophysiologies and accumulated stressors affecting health in evolving urban environments with a strong emphasis on basic disease mechanisms, clinical translation and community health impact.

Ten faculty have been recruited to date as part of the IBio initiative and we are now entering the next phase of thematic-based, programmatic growth. Faculty recruitment (tenured, tenure-track, or research-track) will focus on six primary thematic areas: Behavioral Health (#043728); Bio & Systems Engineering (#043729); Environmental Sciences and Health (#043725); Health Disparities (#043726); Metabolic Diseases (Cardiovascular, Diabetes and Obesity) (#043727); and Translational Neurosciences (#043730). Each theme includes basic discovery-driven research as well as translational, community and implementation sciences cutting across departments, programs, centers, and colleges.

Faculty recruits (tenured or tenure-track) will integrate with departments and colleges or schools consistent with their areas of expertise and shared interests and engage in all aspects of our academic mission including research, education and service. Faculty are expected to either already have established extramural research funding and/or are on a clear path to secure and sustain extramural funding in support of their research programs.

Candidates must have a Ph.D., M.D., Pharm.D. and/or related degree(s) in disciplines aligning with the focus areas and possess a demonstrated track record of exceptional science, creative discovery and/or knowledge translation and application. We would be pleased to receive applications from groups of faculty from one or several institutions who may wish to work together. Qualified candidates should submit applications to the specific thematic position posting # identified above through the Wayne State University Online Hiring System **https://jobs.wayne.edu/applicants/jsp/shared/Welcome_css.jsp**. Applications should include a *curriculum vitae* and a brief narrative describing their research and how it relates to the Integrated Biosciences initiative (http://www.IBio.wayne.edu) with a cover letter addressed to the IBio Steering Committee Chair, Stephen M. Lanier, Ph.D. Vice President for Research. Review of applications for the next phase of recruitments will begin immediately with applications accepted through October 30, 2018. Competitive recruitment packages are available with salary and rank based on qualifications.

Wayne State University, which holds the highest Carnegie Foundation for the Advancement of Teaching designations in both research and community engagement, is a premier, public, urban, comprehensive research university located in the heart of Detroit where students from all backgrounds are offered a rich, high quality education. Our deep-rooted commitment to excellence, collaboration, integrity, diversity and inclusion creates exceptional opportunities for students and faculty in a diverse, global society. WSU encourages applications from women, people of color, and other underrepresented people. Wayne State is an Affirmative Action/Equal Opportunity Employer.

Founded in 1868, Wayne State University offers more than 370 academic programs through 13 schools and colleges to nearly 28,000 students. The campus in Midtown Detroit comprises 100 buildings over 200 acres including the School of Medicine, the Eugene Applebaum College of Pharmacy and Health Sciences and the College of Nursing. The university is home to the Perinatology Research Branch of the National Institutes of Health, the Karmanos Cancer Center, a National Cancer Institute-designated comprehensive cancer center, and a National Institute of Environmental Health Sciences Core Center - *Center for Urban Responses to Environmental Stressors (CURES).*

Yale SCHOOL OF MEDICINE

Department of Neuroscience
New Haven, CT 06520-8001
http://medicine.yale.edu/neuroscience/index.aspx
NEUROSCIENCE FACULTY POSITIONS

The Department of Neuroscience at Yale University seeks to hire faculty in any area of neuroscience, with a preference for candidates who use neuronal or systems level analysis to investigate circuits, behavior or cognition in health and disease.

Emphasis will be placed on recruiting at the level of Assistant Professor, but excellent applicants at Associate Professor level will also be considered. We seek candidates with an exceptional track record, potential for outstanding future achievements, and a wish to participate in a dynamic and recently expanded neuroscience community at Yale that includes the Kavli Institute for Neuroscience, the Program in Cellular Neuroscience, Neurodegeneration and Repair (CNNR) and the Swartz Program in Theoretical Neuroscience. We are especially interested in candidates who will contribute to the diversity of our academic community. Candidates will be supported by a generous start-up and ongoing salary support, and are expected to develop a productive and innovative research program that will include the opportunity to participate in graduate and medical education.

Candidates must hold a Ph.D., M.D., or equivalent degree. Please send a cover letter, curriculum vitae, up to 3 representative publications, a research plan (strictly limited to 2 pages), and arrange for submission of 3 letters of recommendation. All application materials should be submitted electronically through **apply.interfolio.com/54771**. Applications will be reviewed as they are received and until the positions are filled with priority given to those applications received by **November 17, 2018**.

Yale University is an Affirmative Action/Equal Opportunity Employer. Yale values diversity among its students, staff, and faculty and strongly welcomes applications from women, persons with disabilities, protected veterans, and underrepresented minorities.

UT Southwestern Medical Center

TENURE-TRACK POSITIONS

The Department of Physiology invites outstanding scientists with Ph.D., M.D., or equivalent degrees to apply for tenure-track faculty positions at the level of Assistant Professor. Candidates who bring innovative approaches to the study of any under-explored/unexplored questions broadly related to physiology are encouraged to apply. The scientific excellence of the candidates is more important than the specific area of research. These positions are part of the continuing growth of the Department at one of the country's leading academic medical centers. They will be supported by significant laboratory space, competitive salaries, state-of-the-art core facilities and exceptional start-up packages. The University of Texas Southwestern Medical Center is the scientific home to six Nobel Prize laureates and many members of the National Academy of Sciences and Institute of Medicine. UT Southwestern conducts more than 3,500 research projects annually totaling more than $417 million. Additional information about the Department of Physiology can be found at http://www.utsouthwestern.edu/education/medical-school/departments/physiology/index.html.

Applicants should submit a CV, a brief statement of current and proposed research, and a summary of your two most significant publications describing the importance of the work (100-150 words each). Please arrange to have three letters of recommendation sent on his/her behalf. All items should be submitted to: **http://academicjobsonline.org/ajo/jobs/11597**. Completed applications will be reviewed starting **November 1, 2018**. You may email questions to ron.doris@utsouthwestern.edu.

UT Southwestern Medical Center is an Equal Opportunity/Affirmative Action Employer. Women, minorities, veterans and individuals with disabilities are encouraged to apply.

Faculty Positions – Bio & Systems Engineering

Wayne State University (WSU) is recruiting up to15 faculty (open rank) for research and development programs as part of the continued expansion of a broad institutional initiative in Integrative Biosciences. This initiative leverages a new 200,000 sq. ft. Integrative Biosciences Center (IBio) that houses coordinated inter- and trans-disciplinary research teams, and a Clinical Research Center. Programmatic themes include a focus on pathophysiologies and accumulated stressors affecting health in evolving urban environments with a strong emphasis on basic disease mechanisms, clinical translation, and community health impact. Candidates interested in the **Bio & Systems Engineering** thematic area are expected to apply engineering and systems approaches to the focus areas of metabolic diseases, translational neuroscience, environmental health, and health disparities and we plan to recruit up to 3 faculty in this specific thematic area. These approaches include but are not limited to the following:

- Biological and environmental interactions of materials
- Biosensors
- Bionanoengineering
- Neuroprosthetics
- Biosystems engineering

Successful applicants will benefit from a highly collaborative research environment working with colleagues within several colleges at the University. Successful faculty candidates must have a Ph.D. or equivalent degree in biological sciences and engineering relevant to the thematic area, with evidence of peer recognition in the field, a commitment to excellence in research education and training, and the ability to engage with broader IBio disease science themes for the purpose of achieving transformative and translational research gains. Applicants are expected to have already established extramural research funding or be on a clear path to secure and sustain extramural funding in support of their research programs. Faculty recruits will integrate with departments and colleges or schools consistent with their areas of expertise and shared interests and engage in all aspects of our academic mission, including research, education, and service. We would be pleased to receive applications from groups of faculty from one or several institutions who may wish to work together.

Competitive recruitment packages are available with salary and faculty rank based on qualifications. Applicants are encouraged to apply to posting #043729 through the WSU Online Hiring System https://jobs.wayne.edu. Applications will be accepted until positions are filled, but for full consideration this fall, application materials should be submitted by **November 30, 2018**. Applications should include a *curriculum vitae*, a detailed description of research and teaching plans, reprints of two selected publications, the names and addresses of at least three references, and a cover letter addressed to the Vice President for Research indicating the applicant's potential for research synergy within the broader institutional initiative in integrative biosciences.

Wayne State University is a premier, public, urban research university located in the heart of Detroit where students from all backgrounds are offered a rich, high quality education. Our deep-rooted commitment to excellence, collaboration, integrity, diversity and inclusion creates exceptional educational opportunities preparing students for success in a diverse, global society. WSU encourages applications from women, people of color and other underrepresented people. WSU is an affirmative action/equal opportunity employer.

Founded in 1868, Wayne State University offers a range of academic programs through 13 schools and colleges to nearly 28,000 students. The campus in Midtown Detroit comprises 100 buildings over 200 acres including the School of Medicine, the College of Engineering, the Eugene Applebaum College of Pharmacy and Health Sciences and the College of Nursing. The university is home to the Perinatology Research Branch of the National Institutes of Health, the Karmanos Cancer Center, a National Cancer Institute-designated comprehensive cancer center, and a National Institute of Environmental Health Sciences Core Center - *Center for Urban Responses to Environmental Stressors (CURES)*.

Faculty positions in Translational Neurosciences

As part of a broad institutional initiative in Integrative Biosciences, the Translational Neuroscience Initiative (TNI) at Wayne State University (WSU) is recruiting up to three faculty positions (tenured or tenure-track, open rank). The program in translational neuroscience fosters interdisciplinary, integrative, and collaborative approaches to understand basic neuro- and myelin biology and their applications in neurological/psychiatric disorders. Primary areas for recruitment are as follows:

- Myelin development, maintenance, demyelination and remyelination as mechanisms underlying peripheral and central nerve diseases (multiple sclerosis, motor neuron disease, autism, Parkinson's disease)
- Development of new therapies against neurological disorders through cell- or viral vector- based delivery
- Neurogenetics with a focus on neural behavior, pain sensation and/or peripheral and central nerve diseases

The TNI serves as headquarters to promote integrative and collaborative research in basic and clinical neuroscience and is situated in the new 200,000 sq. ft. Integrative Biosciences Center (IBio) that houses coordinated inter- and trans-disciplinary research teams, and a Clinical Research Center. IBio is a fulcrum for leading-edge technology platforms and specialized resources in support of advanced studies in precision medicine. Through research, clinical care, community engagement, and education, the TNI team of researchers and community partners seek to discover, investigate, and solve complex health problems that affect the human nervous system. Successful faculty candidates will have a Ph.D., M.D., or equivalent degree in biomedical sciences relevant to neurosciences with evidence of peer recognition in the field, a commitment to excellence in research education and training, and the ability to engage with broader neuroscience themes for the purpose of achieving transformative and translational research gains. Applicants are expected to have already established extramural research funding or to be on a clear path to secure extramural funding in support of their research programs. Faculty recruits will integrate with departments and colleges or schools consistent with their areas of expertise and shared interests and engage in all aspects of our academic mission, including research, training, instruction and service.

Competitive recruitment packages are available with salary and faculty rank based on qualifications. Applicants are encouraged to apply to posting #043730 through the WSU Online Hiring System https://jobs.wayne.edu. Applications will be accepted until positions are filled, but for full consideration this fall, application materials should be submitted by **November 30, 2018**. Applications should include a *curriculum vitae* and a brief narrative cover letter addressed to the Scientific Director of the TNI and the Vice President for Research, indicating the applicant's potential for research synergy within the TNI and the broader institutional initiative in integrative biosciences.

Wayne State University is a premier, public, urban research university located in the heart of Detroit where students from all backgrounds are offered a rich, high quality education. Our deep-rooted commitment to excellence, collaboration, integrity, diversity and inclusion creates exceptional educational opportunities preparing students for success in a diverse, global society. WSU encourages applications from women, people of color and other underrepresented people. WSU is an affirmative action/equal opportunity employer.

Founded in 1868, Wayne State University offers a range of academic programs through 13 schools and colleges to nearly 28,000 students. The campus in Midtown Detroit comprises 100 buildings over 200 acres including the School of Medicine, the Eugene Applebaum College of Pharmacy and Health Sciences and the College of Nursing. The university is home to the Perinatology Research Branch of the National Institutes of Health, the Karmanos Cancer Center, a National Cancer Institute-designated comprehensive cancer center, and a National Institute of Environmental Health Sciences Core Center - *Center for Urban Responses to Environmental Stressors (CURES)*.

Faculty Postions – Behavioral Health and Disparities

Wayne State University (WSU) is recruiting up to 3 faculty with a focus on behavioral health and disparities (open rank) for research and development programs as part of the continued expansion of a broad institutional initiative in Integrative Biosciences. This initiative leverages a new 200,000 sq. ft. Integrative Biosciences Center (IBio) that houses coordinated inter- and trans-disciplinary research teams, and a Clinical Research Center. Overall programmatic themes of the Integrative Biosciences initiative include a focus on pathophysiologies and accumulated stressors affecting health in evolving urban environments with a strong emphasis on disease mechanisms, clinical translation and community health impact.

Successful candidates in the area of Health Disparities in Behavioral Health will be those who engage in interdisciplinary, behavioral research that addresses persistent disparities across racial, ethnic, gender, socioeconomic, sexual minority, and/ or disability status that will lead to improvements in health. Areas of particular interest include research focusing on lifestyle behaviors (diet, exercise, smoking, substance use) that affect the development or progression of chronic health conditions, translation of basic research findings into effective behavioral interventions, implementation of evidence-based interventions in community settings, and health policy and systems research. Candidates who have experience or interest in establishing partnerships with community stakeholders to develop and refine future behavioral research priorities for minority and disadvantaged populations are preferred.

Faculty recruits (tenured or tenure-track) will integrate with departments and colleges or schools consistent with their areas of expertise and shared interests and engage in all aspects of our academic mission including research, education and service. Faculty are expected to either already have established extramural research funding and/or are on a clear path to secure and sustain extramural funding in support of their research programs.

Candidates must have a Ph.D., M.D., and/or related degree(s) in disciplines aligning with the focus areas and possess a demonstrated track record of exceptional science, creative discovery and/or knowledge translation and application. Qualified candidates should submit applications to the Behavioral Health or Health Disparities posting (posting #043728, #043726) available through the WSU Online Hiring System https://jobs.wayne.edu. Applications should include a *curriculum vitae* and a brief narrative describing their research and how it relates to the Integrated Biosciences initiative (http://www.IBio.wayne.edu) with a cover letter addressed to the program coordinators Drs. Hayley Thompson and Deborah Ellis and the Assistant VP for Integrative Biosciences. Applications will be accepted until positions are filled, but for full consideration this fall, application materials should be submitted by **November 30, 2018**. Competitive recruitment packages are available with salary and rank based on qualifications.

Wayne State University is a premier, public, urban, comprehensive research university (R1 – Carnegie classification) located in the heart of Detroit where students from all backgrounds are offered a rich, high quality education. Our deep-rooted commitment to excellence, collaboration, integrity, diversity and inclusion creates exceptional opportunities for students and faculty in a diverse, global society. WSU encourages applications from women, people of color, and other underrepresented people. WSU is an affirmative action/equal opportunity employer.

Founded in 1868, Wayne State University offers a range of academic programs through 13 schools and colleges to nearly 28,000 students. The campus in Midtown Detroit comprises 100 buildings over 200 acres including the School of Medicine, the Eugene Applebaum College of Pharmacy and Health Sciences and the College of Nursing. The university is home to the Perinatology Research Branch of the National Institutes of Health, the Karmanos Cancer Center, a National Cancer Institute-designated comprehensive cancer center, and a National Institute of Environmental Health Sciences Core Center - *Center for Urban Responses to Environmental Stressors (CURES)*.

Faculty Positions – Diabetes, Obesity and Cardiovascular

As part of a broad institutional initiative within the integrative biosciences and the health science areas, Wayne State University (WSU) is recruiting up to 3 individuals for faculty positions (tenured or tenure-track, open rank) involving the molecular and physiological interplay across **Diabetes, Obesity, and Cardiovascular disease** from both a basic and clinical perspective. Research performed at several departments, centers and institutes at the university foster an interdisciplinary, integrative, and collaborative approach to this area of interest. Building upon existing strengths in mitochondrial biology, cardiac function, metabolic signaling and biosensors, this recruitment targets individuals focused in the following topic areas:

• Cellular mechanisms in obesity and diabetes   • Integrative physiology of metabolic diseases

WSU has a long history of research and clinical care programs in these target areas spanning multiple disciplines including programs in Integrative and Metabolic Research, the Center for Molecular Medicine and Genetics, the Departments of Internal Medicine, Ophthalmology, Visual and Anatomical Sciences, Pathology, Physiology, and Family Medicine and Public Health Sciences in the School of Medicine, as well as in various departments in the College of Pharmacy and Health Sciences.

Successful applicants will benefit from a highly collaborative research environment working with colleagues within several colleges at the University. Successful faculty candidates must have a Ph.D., M.D., or equivalent degree in biomedical sciences relevant to metabolic diseases, with evidence of peer recognition in the field, a commitment to excellence in research education and training, and the ability to engage with broader metabolic disease science themes for the purpose of achieving transformative and translational research gains. Applicants are expected to have already established extramural research funding or be on a clear path to secure and sustain extramural funding in support of their research programs. Faculty recruits will integrate with departments and colleges or schools consistent with their areas of expertise and shared interests and engage in all aspects of our academic mission, including research, education and service.

Competitive recruitment packages are available with salary and faculty rank based on qualifications. Applicants are encouraged to apply to posting #043727 through the WSU Online Hiring System https://jobs.wayne.edu. Applications will be accepted until positions are filled, but for full consideration this fall, application materials should be submitted by **November 30, 2018**. Applications should include a *curriculum vitae* and a brief cover letter to the Assistant Vice President for Integrative Biosciences indicating the applicant's potential for synergy in the targeted thematic areas.

Wayne State University is a premier, public, urban research university located in the heart of Detroit where students from all backgrounds are offered a rich, high quality education. Our deep-rooted commitment to excellence, collaboration, integrity, diversity and inclusion creates exceptional educational opportunities preparing students for success in a diverse, global society. WSU encourages applications from women, people of color and other underrepresented people. WSU is an affirmative action/equal opportunity employer.

Founded in 1868, Wayne State University offers a range of academic programs through 13 schools and colleges to nearly 28,000 students. The campus in Midtown Detroit comprises 100 buildings over 200 acres including the School of Medicine, the College of Engineering, the Eugene Applebaum College of Pharmacy and Health Sciences and the College of Nursing. The university is home to the Perinatology Research Branch of the National Institutes of Health, the Karmanos Cancer Center, a National Cancer Institute-designated comprehensive cancer center, and a National Institute of Environmental Health Sciences Core Center - *Center for Urban Responses to Environmental Stressors (CURES)*.

Assistant/Associate/Full Professor of Chemistry

The Department of Chemistry and Biochemistry at Auburn University invites applications for a tenure-track or tenured, nine-month position with open rank. Auburn University is an institution that is both highly research-active and committed to maintaining teaching excellence as one of the nation's premier land, sea, and space grant institutions. With this position, the department is looking to fill research needs in analytical, physical, or biophysical measurement science that will directly complement existing research in the area of chemical biology. The successful candidate is expected to develop a vigorous, externally funded research program with particular emphasis on analysis of biological systems using methodology such as optical spectroscopy, mass spectrometry, imaging, separations, microfluidics, electrochemistry, biosensors, etc.

Duties also include teaching at both the undergraduate and graduate levels, thus excellent written and interpersonal communication skills are required. A Ph.D. in chemistry, biochemistry, or a closely related field and at least one year of postdoctoral experience are required at the time employment begins. The selected candidate must also meet eligibility requirements to work in the U. S. on the date of appointment (August 2019) and must be able to continue working legally for the proposed term of employment.

All applicants should submit a curriculum vitae, a detailed statement of research plans, a two-page statement of teaching philosophy, and the names and contact information for three professional references. For more information about the College of Sciences and Mathematics and the Department of Chemistry and Biochemistry, please refer to our website: http://www.auburn.edu/chemistry. Review of applications will begin on **October 31, 2018** and continue until the position is filled.

To apply for this position, please visit: **http://aufacultypositions. peopleadmin.com/postings/3079**.

Auburn University is an EEO/Vet/Disability Employer.

Caltech Division of Chemistry & Chemical Engineering

Assistant Professor of Biochemistry
Division of Chemistry and Chemical Engineering
California Institute of Technology
https://applications.caltech.edu/job/bmb

The Division of Chemistry and Chemical Engineering at the California Institute of Technology invites applications for a tenure-track faculty position at the assistant professor level in biochemistry and related areas. Exceptionally well-qualified applicants at the tenured level may also be considered. Candidates with strong commitments to research and teaching excellence are encouraged to apply. The term of the initial appointment is four years and is contingent upon completion of all requirements for a Ph.D. in biochemistry or closely related fields.

Interested candidates should apply electronically at **https://applications. caltech.edu/job/ bmb**. Candidates unable to apply electronically may submit curriculum vita, publication list, teaching statement, description of proposed research, and three letters of recommendation to: Chair of the Biochemistry Search Committee, Division of Chemistry and Chemical Engineering, M/C 164-30, California Institute of Technology, Pasadena, CA 91125.

Applications should be received by **November 15, 2018**.

We are an Equal Opportunity Employer and all qualified applicants will receive consideration for employment without regard to race, color, religion, sex, sexual orientation, gender identity, or national origin, disability status, protected veteran status, or any other characteristic protected by law.

Division of Chemistry and Chemical Engineering | California Institute of Technology, 1200 E. California Boulevard | Pasadena CA 91125 | Phone (626) 395-6099 | Fax (626) 395-6948 | www.cce.caltech.edu

---

Cleveland Clinic
Lerner Research Institute
CENTER FOR MICROBIOME AND HUMAN HEALTH
MULTIPLE FACULTY POSITIONS

The overarching goal of the Cleveland Clinic Center for Microbiome and Human Health (http://cmhh.ccf.org) is to elucidate the role of microbiota in human health and disease, and then exploit this knowledge to develop microbiome-focused therapeutic strategies. We are seeking highly qualified investigators who use definitive approaches to understand the nature of the microbiome and its impact on host physiology and disease susceptibility. Basic, translational, and clinical investigators who are leaders in the field of microbiome-related research are sought to contribute to this mission. In addition, highly promising early career investigators with microbiome-related research interests are encouraged to apply. Candidates are sought with expertise in areas related to microbiology, including, but not limited to, microbe-microbe interactions, microbe-host interactions, microbial ecology, microbial enzymology and biochemistry, bacteriophage therapy, and the role of microbiota in cancer biology and response to immunotherapy.

Individuals will receive faculty appointments at the appropriate level (Assistant, Associate, or Full) in the appropriate basic science or clinical Departments and Institutes. New recruits will join a highly collaborative environment with outstanding opportunities for interactions with basic and translational investigators, as well as with clinicians from multiple specialties. Outstanding facilities, generous start-up funds, and ongoing operational support will be provided.

The Lerner Research Institute, with nearly 175 independent investigators in 8 departments and an annual budget of > $260 million, has a commitment to excellence in basic, translational and clinical research, with an emphasis on interactive, collaborative research. Faculty typically also become members of the Molecular Medicine Graduate Training Program and other programs at Case Western Reserve University School of Medicine, providing access to excellent Ph.D. students. Candidates should submit a complete curriculum vitae and brief statement of research interests by email to: Ms. Paxton Hatcher, 216-445-1937, **hatchep@ccf.org**

Cleveland Clinic is pleased to be an Equal Opportunity Employment/ Affirmative Action Employer: Women/Minorities/Veterans/Individuals with Disabilities. Smoke/drug free environment.

THE HORMEL INSTITUTE
UNIVERSITY OF MINNESOTA

Open Rank Faculty Positions – Cancer Research

The Hormel Institute is a biomedical research center of the University of Minnesota engaged in world-class medical research. Researchers have complete access to an array of state-of-the-art equipment that includes a cryo-electron microscope (Titan Krios & Tecnai G2 Spirit Biotwin); FACS cell sorter; protein crystallography robotics & defraction system; nano-HPLC-AB SCIEX triple TOF 5600 mass spectrometry; a Blue Gene/L & 5 racks of Linux GPU supercomputers for computational biology & bioinformatics. Currently accepting applications for 3-5 new faculty positions at the levels of Assistant, Associate, or Full Professor.

Faculty at The Hormel Institute focus 100% on research. Lab Section Leaders establish & manage a lab, hire & supervise lab faculty/staff, maintain sufficient funding for the lab's operation, foster relationships with internal & external collaborators, publish scientific results, & present at relevant events & conferences.

Qualifications: Candidates must demonstrate the ability to establish an independent, extramurally funded program of research that will complement ongoing programs focusing on various areas of cancer research. Ph.D. & 2-3 years postdoc experience is required for Assistant Professor & external funding is preferred. Proven ability to obtain funding is required for Associate & Full Professor ranks. Preferred: Expertise or current projects involving **cryo-electron microscopy**, translational, physician scientists (MD/PhD), **epidemiology** & populations studies.

To apply, please submit a single PDF containing CV, Research Interests, and 3 Letters of Reference to **bearl@umn.edu**. Please include the words "Faculty Search" in the subject line.

Faculty Positions in Chemistry and Chemical Biology

RUTGERS
School of Arts and Sciences

The Department of Chemistry and Chemical Biology (CCB) at Rutgers, The State University of New Jersey is seeking faculty candidates for three tenure-track appointments at the assistant or associate level at our New Brunswick/Piscataway location. Successful candidates and their research groups will be housed in our brand new 145,000 sf, state-of-the-art chemistry building.

We especially welcome applicants with interests in (1) theoretical and computational chemistry; (2) biophysical chemistry with a focus on mass spectrometry, biomolecular design and/or spectroscopic approaches to biological systems; (3) organic, inorganic or physical chemistry focusing on the molecular basis of aging or neurodegenerative diseases – including redox chemistry, proteomic and other approaches.

CCB faculty are expected to demonstrate an ability to build and sustain a world-class research program. They must have a significant record of scholarship in fundamental research, and are expected to contribute to CCB's teaching and service missions.

How to Apply: Applicants should submit the following to **http://jobs.rutgers.edu/postings/74621** and indicate one or more preferred search focus areas. Please include (1) cover letter and detailed CV; (2) four to six-page research development statement; (3) one-page teaching statement; 4) Contact information for three references. Questions about this search should be directed to chemchair@chem.rutgers.edu. Review of applications will begin on **October 15, 2018** and will continue until the position is filled.

All qualified applicants will receive consideration for employment without regard to race, color, religion, sex, sexual orientation, gender identity or expression, national origin, disability, protected veteran status or any other classification protected by law. All job offers are contingent upon successful pre-employment background screening prior to commencement of employment.

UNIVERSITY of ROCHESTER
MEDICAL CENTER

Assistant/Associate Professor
Center for Biomedical Informatics (CBI) and
Wilmot Cancer Institute (WCI)

The University of Rochester Medical Center (URMC) is expanding its research activities in the area of biomedical informatics. We are seeking investigators at the Assistant or Associate Professor level with independent and collaborative research programs. Six to ten faculty will be recruited into tenure track positions over the next three years.

Here we are searching for investigators with expertise in the broad areas of bioinformatics, quantitative or systems biology and a research program in cancer biology. Successful candidates will hold appointments in the CBI and the WCI. Academic department affiliation will be determined according to the best fit. Candidates with research experience in the following areas are highly encouraged to apply: Genome and transcriptome informatics, systems biology, population genomics, machine learning, data mining, computational modeling, multi-dimensional data integration of clinical, genomic, metabolome, gene expression or epigenetic data.

URMC offers attractive start-up packages and has a strong commitment to career development. The successful candidate is expected to develop a competitive research program, attract external funding, and participate in graduate education. Major recent institutional investments have created an outstanding research data–driven infrastructure, exemplified by the Health Sciences Center for Computational Innovation, and the recent $50 million University investment to build the Institute for Data Science.

Applicants should submit a letter of application, CV, statement of research interests/plans, and arrange to have three letters of recommendation sent to: **anne_reed@urmc.rochester.edu**. Inquiries can be directed to Dirk Bohmann (dirk_bohmann@urmc.rochester.edu) or Hartmut Land (land@urmc.rochester.edu).

The University of Rochester is an Equal Opportunity Employer and has a strong commitment to diversity; it actively encourages applications from groups underrepresented in higher education.

Department of Genetics
Yale University School of Medicine

Yale UNIVERSITY

Professor of Genetics and
Director of the Center for Genomic Health

The Department of Genetics at Yale University School of Medicine is searching for a Professor of Genetics with an outstanding record of transformative scientific achievements in Human Genetics and Genomics. We expect that the candidate will lead a vigorous cross-disciplinary research program focused on identifying and characterizing genetic drivers of human disease. As a leader of human genetics both within the department and across the Yale School of Medicine, the successful candidate will have the opportunity to recruit other human geneticists to the Genetics Department and lead a new program in precision medicine as the Scientific Director of the Yale Center for Personalized Medicine and Genomic Health.

We are looking for a dynamic, internationally recognized scientist (Ph.D., M.D., or M.D./Ph.D.) with an outstanding research record of scientific discoveries, as well as a strong track record of training the best innovators in the field of human genetics and genomics.

To apply, please submit your CV to **http://apply.interfolio.com/45539** to the attention of **Antonio Giraldez**, Chair of Genetics. Applications will be reviewed until the position is filled. Inquiries should be addressed to **neltja.brewster@yale.edu**.

Tenure-Track Faculty Position
(Assistant/Associate/Full Professor)

The Department of Genetics at Yale University School of Medicine invites applications for junior or senior tenure-track faculty positions. The search is open to investigators from all areas of biological and biomedical research. We are particularly interested in applicants working in one of the following areas: Developmental Biology, Imaging, Quantitative Biology, Computational Biology, Genomics, Systems Biology, and Genetics. Applications from investigators working at the interface of these areas will be strongly considered. The rank of the appointment will be commensurate with experience and the positions come with a substantial start-up package.

The Department of Genetics comprises an exceptional group of 31 primary basic science faculty with research interests including fundamental aspects of Developmental Biology, Genetics, Genomics and Epigenetics, using different model systems including flies, worms, fish and mouse, and humans (https://medicine.yale.edu/genetics/). The Department is closely associated with science initiatives at Yale including The Cancer Center, The Center for Neuroscience, The Stem Cell Center and the Yale Center for Genome Analysis.

Candidate must hold a Ph.D., M.D., or equivalent degree. Applicants should upload a cover letter, a curriculum vitae, a description of previous research (1 page), a concise statement of research plans (up to 2 pages), reprints of 2 publications, and the names of 3 references to the Interfolio website at: **https://apply.interfolio.com/52614**. Specific inquiries about the position may be sent to the attention of **Dr. Antonio Giraldez**, Chair of the Department of Genetics, at genetics.admin@yale.edu. Applications will begin to be evaluated on **November 1, 2018**.

Interviews will take place as part of a multidisciplinary symposium including candidates for different searches. Please reserve the dates of January 15, January 22 and February 11 (snow date) as potential dates for the symposium in case you are selected for an interview.

Yale University is an Affirmative Action/Equal Opportunity Employer. Yale values diversity among its students, staff, and faculty and strongly welcomes applications from women, persons with disabilities, protected veterans, and underrepresented minorities.

Assistant Scientist Cancer Biology Group, Sanford Research

The Cancer Biology Group at Sanford Research invites applications for full-time faculty at the rank of Assistant Scientist within Sanford Research in Sioux Falls, SD, with commensurate rank of Assistant Professor at the Sanford School of Medicine at the University of South Dakota. Sanford Research is the non-profit research branch under Sanford Health.

We seek outstanding scientists with research programs that span all areas of cancer research, especially those relevant to cancer immunology or immunotherapy. The successful candidate will have an opportunity to become a project leader on the NIH-funded Cancer Biology CoBRE. Significant institutional support, including modern laboratory space and state-of-the-art facilities will be provided. A comprehensive benefits package will be tailored to the candidate's qualifications.

Qualifications
Applicants should hold a PhD, MD or MD/PhD degree and complement the existing strengths and interdisciplinary and collaborative nature of Sanford Research. Physician Scientists are encouraged to apply. Candidates will be expected to develop independent research programs and secure extramural funding.

Application
Sanford Health is an Equal Opportunity/Affirmative Action Employer. Applicants should submit a single PDF that includes: 1) detailed curriculum vitae, 2) description of research experience and future research plans with details on relevance of their research to cancer biology and/or cancer immunology, and 3) three letters of recommendation. If any of the information above is missing, the submission will not be considered. Submit materials via email to: researchrecruitment@sanfordhealth.org

SANFORD HEALTH

019036-00762. 8/18

Associate Scientist/Scientist Population Health Group, Sanford Research

The Population Health Group at Sanford Research invites applications for full-time faculty at the rank of Associate Scientist or Scientist within Sanford Research in Sioux Falls, SD, with commensurate rank of Associate Professor or Professor at the Sanford School of Medicine at the University of South Dakota. Sanford Research is the non-profit research branch under Sanford Health.

The Population Health Group is comprised of investigators focused on public health, epidemiology, and population health with a strong emphasis on American Indian and rural population health. Ideal candidates will have an existing research portfolio in population health, health disparities, or a similar field of study. Additionally, the candidate will become the Principal Investigator of a currently funded NIH Center for Biomedical Research Excellence (CoBRE) grant. The candidate will mentor junior faculty, oversee relevant scientific cores, and utilize the CoBRE and institutional resources to grow and sustain the Center, including recruitment of several new investigators during the tenure of the grant. This candidate would also have the opportunity to engage in translational health services research leveraging Sanford's integrated data warehouse (clinical and claims data) to inform healthcare delivery. Significant institutional support, including modern laboratory space and state-of-the-art facilities will be provided. A comprehensive benefits package will be tailored to the candidate's qualifications.

Qualifications
Applicants should hold a PhD, MD or MD/PhD degree and complement the existing strengths and interdisciplinary and collaborative nature of Sanford Research. Physician Scientists are encouraged to apply. Ideal candidates will have a strong record of independent investigator-initiated grant funding and ideally program grant funding in population health (or a similar field), managing large budgets, mentoring early-career investigators to funding success, and experience in developing collaborations with various communities and institutional partners.

Application
Sanford Health is an Equal Opportunity/Affirmative Action Employer. Applicants should submit a single PDF that includes: 1) detailed curriculum vitae, 2) description of research experience and future research plans with details on relevance of their research to population health (or related topic), and 3) three letters of recommendation. If any of the information above is missing, the submission will not be considered. Submit materials via email to: researchrecruitment@sanfordhealth.org

SANFORD HEALTH

019036-00774 8/18

Cleveland Clinic
Lerner Research Institute

CENTER FOR MICROBIOME AND HUMAN HEALTH
Microbial Culturing and Engineering Core Director

Our Institution is building a new Center focused on the microbiome, its involvement in human health, and its potential as a target for therapeutic interventions. We are seeking a highly qualified new or experienced investigator who is an expert in microbial culturing to establish a new Core in our Center. The applicant should have hands on experience in the field of microbiome-related research and be familiar with a wide array of methods for microbial culturing and strain manipulation under both aerobic and anaerobic conditions.

The successful candidate will have a broad expertise in strain engineering/recombineering and a respected publication record in the field. Importantly the candidate must have the ability to stay abreast with cutting-edge technology and demonstrate good oral and written English communication, proficiencies that are central to providing customer service and technical oversight. The director will collaborate with institutional researchers interested in areas related to microbiology, including microbe-microbe interactions, microbe-host interactions, microbial enzymology and biochemistry, bacteriophage therapy and gnotobiotics. The core director is expected to meet with basic, translational and clinical investigators and help optimally plan experiments focused on microbiomerelated research. The core director will also organize efforts to provide investigators with conventional or genetically-modified bacterial strains for functional studies. Highly promising early career investigators with microbiome-related research interests are encouraged to apply. Alternatively, established investigators that have an established research program in one of these fields and the ability to direct and instruct team members to carry out core functions will also be favorably considered. Faculty appointments will be made at the appropriate level (Staff Scientist, Assistant, Associate, or Full Staff) based upon candidate credentials. The successful candidate will join a highly collaborative environment with outstanding opportunities for interactions with basic and translational investigators, as well as with clinicians from multiple specialties. Outstanding facilities, generous core start-up funds, and ongoing core operational support will be provided.

The Lerner Research Institute, with nearly 200 independent investigators in 10 departments and an annual budget of >$240 million, has a commitment to excellence in basic, translational and clinical research, with an emphasis on interactive, collaborative research. Faculty are members of the Molecular Medicine Graduate Training Program at Case Western Reserve University School of Medicine, and may join other Programs, providing access to excellent Ph.D. students enrolled in welldeveloped training programs.

Candidates should submit a complete curriculum vitae and brief statement of research interests by email to Ms. Kelsey Shirey: **shireyk@ccf.org**.

The Cleveland Clinic Foundation is an Equal Opportunity/Affirmative Action Employer in a smoke/drug free environment.

JOB FOCUS: FACULTY

Professor (Faculty Rank)
Department of Mechanical Science and Engineering
College of Engineering
University of Illinois Urbana-Champaign

The Department of Mechanical Science and Engineering at the University of Illinois at Urbana-Champaign invites applications for multiple faculty positions at all ranks. Candidates are sought in all technical subdisciplines of mechanical science and engineering including design, thermosciences, solid and fluid mechanics, and dynamics and control, with particular interest in the broad application areas of mechanics of materials, robotics, and cyber-physical systems. Ideal candidates include those who demonstrate evidence of a commitment to diversity, equity, and inclusion through research, teaching, and/or service endeavors.

Qualified senior candidates will also be considered for a named appointment as part of the Grainger Engineering Breakthroughs Initiative, supported by a $100-million gift from the Grainger Foundation.

A doctoral degree is required, all opportunities are full-time, tenure or tenure-track, 9-month appointments paid over 12 months, with salary commensurate with qualifications and experience. Applications received by **November 15, 2018** will receive full consideration. Early applications are strongly encouraged. Interviews may take place before the given date; applications received after November 15, 2018 may be considered until positions are filled. The expected start date of a position offered/accepted through this search is August 16, 2019, but other start dates will be considered.

A full position description and information on how to apply can be found on the University of Illinois at Urbana-Champaign online jobsite **http://jobs.illinois. edu**. For further information regarding application procedures, please address questions to: **mechse-facultyrecruiting@illinois.edu**.

The University of Illinois conducts criminal background checks on all job candidates upon acceptance of a contingent offer.

The U of I is an EEO Employer/Vet/Disabled
www.inclusiveillinois.illinois.edu

I ILLINOIS

NRC Research Associateship Programs

The National Academy of Sciences, Engineering, and Medicine offers postdoctoral and senior research awards on behalf more than 20 U.S. federal research agencies and affiliated institutions with facilities at over 100 locations throughout the U.S. and abroad.

We are actively seeking highly qualified candidates including recent doctoral recipients and senior researchers. Applications are accepted during four annual review cycles (with deadlines of November 1, February 1, May 1, and August 1,).

Awardees have the opportunity to:
- conduct independent research in an area compatible with the interests of the sponsoring laboratory
- devote full-time effort to research and publication
- access the excellent and often unique facilities of the federal research enterprise
- collaborate with leading scientists and engineers at the sponsoring laboratories

Benefits of an NRC Research Associateship award include:
- 1 year award, renewable for up to 3 years
- Stipend ranging from $45,000 to $80,000, higher for senior researchers
- Health insurance, relocation benefits, and professional travel allowance

Applicants should hold, or anticipate receiving, an earned doctorate in science or engineering. Degrees from universities abroad should be equivalent in training and research experience to a degree from a U.S. institution. Some awards are open to foreign nationals as well as to U.S. citizens and permanent residents.

The National Academies of Sciences, Engineering, and Medicine's Fellowships Office has conducted the NRC Research Associateship Programs in cooperation with sponsoring federal laboratories and other research organizations approved for participation since 1954. Through national competitions, the Fellowships Office recommends and makes NRC Research Associateship awards to outstanding postdoctoral and senior scientists and engineers for tenure as guest researchers at participating laboratories. A limited number of opportunities are available for support of graduate students in select fields.

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

GROUP LEADER POSITIONS IN STEM CELL & DEVELOPMENTAL BIOLOGY AND ORGANOID RESEARCH

We invite applications for fully funded research group leader positions at IMBA, the Institute of Molecular Biotechnology of the Austrian Academy of Sciences in Vienna.

IMBA is one of Europe's leading life sciences institutes. Together with the IMP, GMI and MFPL, IMBA is part of the Vienna BioCenter, a vibrant research campus home to 1400 scientists and 700 students from over 40 different countries. IMBA group leaders are part of a dynamic and collaborative faculty, are generously funded, and recruit students via the highly competitive international Vienna BioCenter PhD program.

The recruitment package includes a competitive salary, four fully funded positions (two PhD students, one postdoctoral researcher, one technician), and generous consumables & investment budgets. Research groups enjoy free access to a broad portfolio of scientific core facilities, including a mouse house, microscopy & cytometry, mass spectrometry, bioinformatics, and stem cell and genome engineering facilities. Successful candidates will be offered an initial contract of five years, which is foreseen to be extended to eight years, subject to external review. IMBA is committed to achieving gender balance and strongly encourages applications from women. Appointment will be based on merit alone.

Vienna is a multicultural capital city, consistently ranking amongst the top three cities worldwide for the quality of living. Subsidized child care facilities are available at the Vienna BioCenter.

For further information, see:
www.imba.oeaw.ac.at and **viennabiocenter.org**

Applications should include a cover letter, a CV, a concise summary of past achievements and future research plans. We further ask for three reference letters. In your publication list please explain briefly the relevance of your most important work.

Applications should be submitted electronically to our job portal **www.imba.oeaw.ac.at/StemCellSearch** by November 15, 2018.

WWW.IMBA.OEAW.AC.AT

DR. BOHR-GASSE 3 | 1030 VIENNA, AUSTRIA
T: +43 1 79044 | F: +43 1 79044-110

IMBA
Institute of Molecular Biotechnology
of the Austrian Academy of Sciences
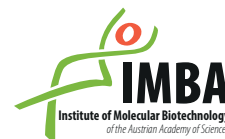
GROUP LEADER POSITIONS IN MOLECULAR, CELLULAR, AND ORGANISMAL BIOLOGY

We invite applications for fully funded research group leader positions at IMBA, the Institute of Molecular Biotechnology of the Austrian Academy of Sciences in Vienna.

IMBA is one of Europe's leading life sciences institutes. Together with the IMP, GMI and MFPL, IMBA is part of the Vienna BioCenter, a vibrant research campus home to 1400 scientists and 700 students from over 40 different countries. IMBA group leaders are part of a dynamic and collaborative faculty, are generously funded, and recruit students via the highly competitive international Vienna BioCenter PhD program.

The recruitment package includes a competitive salary, four fully funded positions (two PhD students, one postdoctoral researcher, one technician), and generous consumables & investment budgets. Research groups enjoy free access to a broad portfolio of scientific core facilities, including a mouse house, microscopy & cytometry, mass spectrometry, bioinformatics, and stem cell and genome engineering facilities. Successful candidates will be offered an initial contract of five years, which is foreseen to be extended to eight years, subject to external review. IMBA is committed to achieving gender balance and strongly encourages applications from women. Appointment will be based on merit alone.

Vienna is a multicultural capital city, consistently ranking amongst the top three cities worldwide for the quality of living. Subsidized child care facilities are available at the Vienna BioCenter.

For further information, see:
www.imba.oeaw.ac.at and **viennabiocenter.org**

Applications should include a cover letter, a CV, a concise summary of past achievements and future research plans. We further ask for three reference letters. In your publication list please explain briefly the relevance of your most important work.

Applications should be submitted electronically to our job portal **www.imba.oeaw.ac.at/GroupLeaderSearch** by December 15, 2018.

WWW.IMBA.OEAW.AC.AT

DR. BOHR-GASSE 3 | 1030 VIENNA, AUSTRIA
T: +43 1 79044 | F: +43 1 79044-110

IMBA
Institute of Molecular Biotechnology
of the Austrian Academy of Sciences
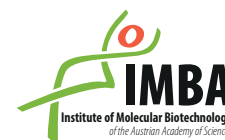
By **Amir Sheikhi**

More than my publications

I should have been celebrating. I had just hit the "submit" button on the online portal for doctoral theses, concluding my Ph.D. studies. And I did feel a sense of satisfaction. I had worked hard and was proud of what I had accomplished. But I was also overwhelmed by depression and anxiety about my future. Almost all my recently graduated friends had several published papers from their Ph.D. work; I had none. I thought my chances at the academic career I so wanted were ruined. I felt lost and alone in an endless desert.

I could point to reasons why I hadn't published. I had changed fields between my master's degree and Ph.D., and it had taken time to get up to speed in my new discipline. I needed to figure out how to work with my new lab's complex custom-built instruments and tune them for my experiments. There were no senior graduate students or postdocs in the lab to train me at the time, which made this task even more challenging. I eventually got my experiments to work, and I learned a tremendous amount along the way, scientifically and personally. But that is not the same as generating publishable results.

As I approached the end of my Ph.D., I was anxious to publish. Yet my Ph.D. scholarship only covered 3 years, with limited support for a fourth year. My time was up. I had enough data to write and defend my thesis, but the results were not ready for publication in a peer-reviewed journal.

So there I was: a newly minted Ph.D. hoping to pursue an academic career but without any papers from my primary project. My applications for postdoctoral research positions were either rejected or ignored. I needed to try something different.

A few years earlier, a grad student friend and I had worked on a side project, which we published with the help of one of his advisers. The work was in some ways very different from my scientific training and my primary Ph.D. project (I was a chemical engineer, and my friend's lab was in the chemistry department), but I had enjoyed applying my knowledge and perspective in this new realm. Now, I thought this experience might help me get a position in that lab.

I arranged a meeting with the lab's principal investigator (PI) and another senior PI. They were working on a

"I shouldn't have defined myself based on my publication record."

joint project, and we had a lively discussion about it and where they were having trouble. I found that my different scientific background and the fundamental knowledge I had developed during my doctoral studies enabled me to propose plausible solutions and alternative approaches that hadn't occurred to them. They offered me a temporary research associate position, with the possibility of staying on longer if things went well. It wasn't the prestigious postdoc position I had initially been looking for, but I was happy to have found some water in my post-Ph.D. desert.

I thrived in this new position. I finished the project in a few months—the PIs had originally expected that it would take a year—and they invited me to continue as an official postdoctoral fellow. I published some of my new work relatively quickly and initiated and completed a number of new projects. I also eventually published the results of my Ph.D. project. Most important, I enjoyed my work. I've since moved on to another position, and despite my earlier publication drought, I am optimistic about my future career prospects.

Looking back, I wish I could tell my recently graduated self that I shouldn't have defined myself based on my publication record. My list of publications doesn't matter as much as the knowledge I gained in the process and the sense of purpose that kept me going. And I wish I could tell everyone else to try to look past publications, at least a little, when evaluating other scientists. Yes, publications are important. But a lack of publications doesn't mean a lack of intelligence, skills, or worth. ∎

Amir Sheikhi is a Canadian Institutes of Health Research postdoctoral fellow at the University of California, Los Angeles. Send your career story to SciCareerEditor@aaas.org.