Bronze Age bonanza: Excavating Britain's Pompeii *p. 210*

How habitat changes threaten butterflies *p. 216*

Ducklings understand same and different *pp. 222 & 286*

# Science

$15
15 JULY 2016
sciencemag.org

AAAS

SPECIAL ISSUE

# NATURAL HAZARDS

# CONTENTS

216
Can we halt butterfly declines?

15 JULY 2016 • VOLUME 353 • ISSUE 6296

## NEWS

### IN BRIEF



### IN DEPTH



### FEATURE



## INSIGHTS

### PERSPECTIVES



PHOTO: CARSTEN PETER/NATIONAL GEOGRAPHIC CREATIVE

SPECIAL SECTION

## NATURAL HAZARDS




**ON THE COVER**

Ash from Mount Sinabung, one of Indonesia's 130 active volcanoes, billows over the Sumatran village of Sukanalu on 13 June 2015. Studies of volcanoes and other natural hazards have yielded better warnings, risk assessments, and disaster response, but dangers remain. A special section beginning on page 230 and online at http://scim.ag/29qQGyL spotlights new tools and ideas for taming nature's fury. *Photo: NurPhoto/NurPhoto via Getty Images*

SEE ALSO ▸ EDITORIAL P. 201 ▸ POLICY FORUM P. 224 ▸ RESEARCH ARTICLE P. 262 ▸ Q&A ▸ PODCAST

Sicily's Mount Etna erupts at night.



### POLICY FORUM

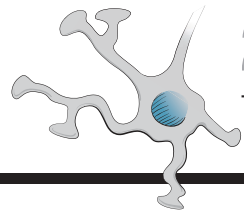






Published by AAAS

# CONTENTS



**218 & 300**
The dynamic neuroepigenome

15 JULY 2016 • VOLUME 353 • ISSUE 6296






**SCIENCE (ISSN 0036-8075) is published weekly on Friday, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005.** Periodicals mail postage (publication No. 484460) paid at Washington, DC, and additional mailing offices.  Domestic individual membership and subscription (51 issues): $165 ($74 allocated to subscription). Domestic institutional subscription (51 issues): $1522; Foreign postage extra: Mexico, Caribbean (surface mail) $55; other countries (air assist delivery) $89. First class, airmail, student, and emeritus rates on request. Canadian rates with GST available upon request, GST #1254 88122. Publications Mail Agreement Number 1069624. **Printed in the U.S.A.**
**Change of address:** Allow 4 weeks, giving old and new addresses and 8-digit account number. **Postmaster:** Send change of address to AAAS, P.O. Box 96178, Washington, DC 20090–6178. **Single-copy sales:** $15.00 current issue, $20.00 back issue prepaid includes surface postage; bulk rates on request. 

Science
AAAS
1200 New York Avenue, NW, Washington, DC 20005
Clarendon House, Clarendon Road, Cambridge, UK CB2 8FH



**SUBSCRIPTION SERVICES** For change of address, missing issues, new orders and renewals, and payment questions: 866-434-AAAS (2227) or 202-326-6417, FAX 202-842-1065. Mailing addresses: AAAS, P.O. Box 96178, Washington, DC 20090-6178 or AAAS Member Services, 1200 New York Avenue, NW, Washington, DC 20005

**INSTITUTIONAL SITE LICENSES** 202-326-6730 **REPRINTS:** Author Inquiries 800-635-7181 **COMMERCIAL INQUIRIES** 803-359-4578 **PERMISSIONS** 202-326-6765, permissions@aaas.org **AAAS Member Services** 202-326-6417 or http://membercentral.aaas.org/discounts



**INFORMATION FOR AUTHORS** See pages 624 and 625 of the 5 February 2016 issue or access www.sciencemag.org/authors/science-information-authors

EDITORIAL

# Hazards without disasters

A natural hazard need not become a human disaster if society learns and applies lessons in preparation and resilience. Earthquake history speaks well to this—engineered structures need to stand up to strong shaking. Chile learned this lesson before its 2010 earthquake of magnitude 8.8. Because it had already enforced seismic provisions of building codes, there was little loss of life due to damage to buildings. Engineered structures also performed very well during the giant 2011 Tohoku earthquake in northeast Japan; however, approximately 20,000 lives were lost to the ensuing tsunami. What survival strategies are available for communities at risk for tsunamis?

One community that has recently acted on tsunami risk is the Ocosta School District of southwest coastal Washington state. It is situated in Cascadia, a region that includes the Pacific coast of southern British Columbia, Washington, Oregon, and northern California. This area has a geological history of great earthquakes and associated tsunamis, generated offshore along the Cascadia Subduction Zone. The most recent of these tsunamis, in January 1700, caused massive flooding and subsidence in the coastal lowlands of North America and a tsunami recorded in Japan. Today, the campus of three Ocosta schools is located on a Washington peninsula that the 1700 tsunami invaded. There are only ~15 min between a Cascadia earthquake and the arrival of the first tsunami wave on this sandy strip. The good news is that last month, a safe haven atop a new elementary school was completed there, North America's first engineered tsunami refuge: a flat roof 30 feet above ground, accessible from each of four heavily reinforced corners by extra-wide stairways, supported further by 169 pilings sunk as much as 50 feet deep that would anchor the structure should a tsunami press against them. The rooftop has space for as many as 2000 tsunami refugees from the school and surrounding residences.


***"The conversation shifted from the...risk...to...taking actions..."***

This outcome offers lessons on how to prevent hazards from becoming disasters. Success at Ocosta depended on partnerships: earthquake geologists, tsunami modelers, social scientists, emergency managers, structural engineers, landscape architects, an elected school board, and the district superintendent. The conversation shifted from the sensational risk itself (the tsunami), to how the community could control its own destiny by taking actions that, in the superintendent's words, put students' well-being and safety above all else. The bond measure that included an additional $2 million for the tsunami refuge was overwhelmingly passed by district voters despite a weak economy and high unemployment. This amounted to 15% of the construction cost of the new school.

The Ocosta example is unusual in its foresight, but it need not be. The safe haven resulted from exceptional local initiative, but the initiative took cues from the 2011 Tohoku disaster and was founded on scientific discoveries about Cascadia's earthquake and tsunami potential. This combination has triggered action before the next Cascadia tsunami, in what the U.S. Federal Emergency Management Agency calls pre-disaster mitigation. One 2000-person refuge is, of course, just a small step toward reducing losses to the next Cascadia tsunami. But research shows that "the strongest motivator of taking preparedness actions is when average people share what they have done to prepare with other individuals who have not done much."* And plans for additional tsunami safe havens are now proceeding elsewhere in Cascadia.

**–Marcia McNutt†**

*Marcia McNutt is president of the U.S. National Academy of Sciences and a former editor-in-chief of* Science. *Email: mmcnutt@nas.edu*

PHOTO: (INSET) REUTERS/MAINICHI SHIMBUN; (TOP RIGHT) STACEY PENTLAND PHOTOGRAPHY

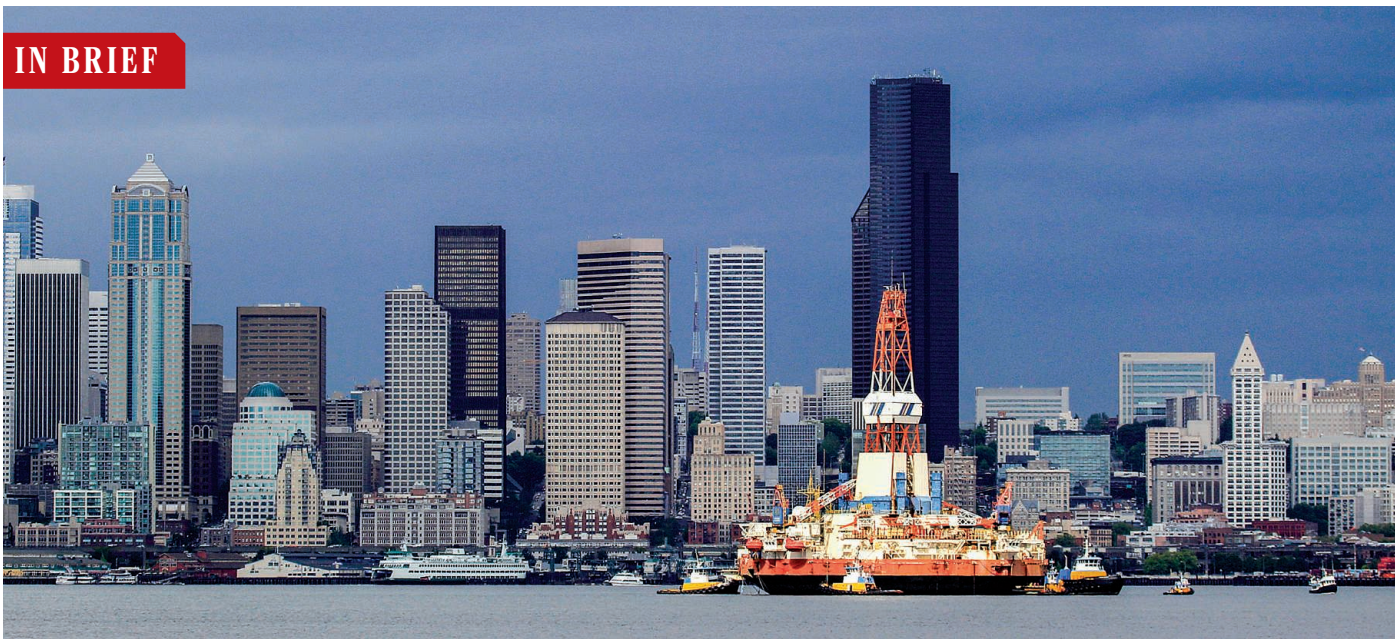
*M. M. Wood *et al., Risk Anal.* **32**, 601 (2012). †M.M. wrote this editorial while editor-in-chief of *Science*.

10.1126/science.aah4247

# NEWS

> “[Mary Todd Lincoln] was simply a woman with a biochemically injured mind struggling in a complicated, relentlessly demanding environment.”
>
> **Cardiologist and Intel Medical Director John Sotos** to *The Washington Post*, on his hypothesis that the former first lady suffered from a vitamin B-12 deficiency.

## IN BRIEF



Royal Dutch Shell’s floating drilling platform *Kulluk*, shown here near Seattle, Washington, in 2011, ran aground in 2012.

# United States sets safety rules for Arctic drilling

The White House on 7 July released the final version of its Arctic safety regulations for future offshore drilling in the Beaufort and Chukchi seas. The new regulations focus specifically on exploratory drilling from floating vessels on the U.S. Outer Continental Shelf (OCS). They build in part on previous Department of the Interior reports after Royal Dutch Shell’s drilling rig, the *Kulluk*, ran aground off the coast of Sitkalidak Island in Alaska in 2012. The regulations now require oil companies to submit a detailed operations plan for their proposed Arctic OCS drilling. Additionally, companies must have containment equipment ready to deploy promptly in case of a blowout; have a second rig available to drill a relief well under harrowing Arctic sea conditions; be able to predict, track, and respond to floating ice conditions and adverse weather; and be able to deploy personnel and equipment quickly in case of an oil spill. Shell had received conditional approval in May 2015 to drill in the Chukchi Sea off Alaska’s coast. The company drilled an exploratory well over the summer, but by October 2015 it abandoned further drilling plans, stating that the well didn’t yield sufficient oil and gas to justify its $7 billion price tag, and citing an “unpredictable” regulatory environment. Environmental groups hailed the new regulations as a step in the right direction, but Alaskan leaders, including the state’s governor, Bill Walker, and Senator Lisa Murkowski (R), say the rules could discourage investment and development in the region.

AROUND THE WORLD

## AIDS epidemic is over, down under

SYDNEY, AUSTRALIA | Australia’s AIDS epidemic is over, and the country could virtually eliminate HIV infections by 2020, according to an announcement by researchers and community experts this week. The near-zero AIDS diagnoses are the culmination of a 33-year response spanning science, medicine, and community outreach that began soon after HIV arrived in Australia. Although effective and accessible therapies mean that few people progress to AIDS, too many people are still becoming infected with HIV, the statement noted. About 1000 Australians are still diagnosed with HIV every year. The groups behind the announcement, which include the Australian Federation of AIDS Organisations, the Kirby Institute for infection and immunity in society, the Peter Doherty Institute, Pacific Friends of the Global Fund, and the National

PHOTO: AP PHOTO/ELAINE THOMPSON





Published by AAAS

Association of People with HIV Australia, called for a long-term strategy to stop HIV from spreading that would target "vulnerable communities" including women, isolated gay men, and non-English speakers. "The AIDS public health threat has morphed into an HIV prevention challenge," says medical epidemiologist Andrew Grulich of the Kirby Institute, which leads Australia's national surveillance of HIV/AIDS and is based at the University of New South Wales in Sydney. **http://scim.ag/AusAIDSend**

## Deaths halt cancer trial

SEATTLE, WASHINGTON | The U.S. Food and Drug Administration (FDA) last week halted a company's clinical trial of a promising cancer treatment after three young adults with leukemia died after receiving it. The therapy, called chimeric antigen receptor (CAR) T cell therapy, involves genetically engineering a patient's T cells, which play a central role in immunity, and then returning them to the body to target the disease. Because of some remarkable successes in very sick patients, it has taken the cancer world by storm and clinical trials are testing the approach on different forms of cancer. Although CAR T cell therapy carries a risk of serious side effects, almost no deaths have been previously reported. The company running the clinical trial, Juno Therapeutics, suggests that the deaths were due to a strategy oncologists have eyed with interest and tried in some other trials: combining CAR T cell therapy with another treatment to bolster its effectiveness. In this case, patients received the chemotherapy drug fludarabine, which aims to wipe out a patient's existing T cells, giving the engineered cells room to multiply. Juno thinks fludarabine was the culprit. FDA will now review documents submitted by the company as it weighs what to do next.

## Senate OKs GM food labeling bill

WASHINGTON, D.C. | The U.S. Senate has passed a bill that would create a national standard for labeling food made with genetically modified organisms (GMOs). The vote marks a win for those who have been pushing the federal government to set a single national standard in hopes of heading off a patchwork of state labeling laws, such as Vermont's law that went into effect on 1 July. The legislation blocks states from issuing their own mandatory labeling laws and requires food manufacturers to inform consumers of the presence of GMOs with one of three different labels—a U.S. Department of Agriculture symbol, a scanner- or smartphone-readable code, or a label stating plainly that GMOs are present. The bill was the result of a deal crafted by senators Pat Roberts (R–KS) and Debbie Stabenow (D–MI), the chair and ranking member, respectively, of the Senate agriculture committee. The Senate bill now goes to the House of Representatives, and food and farm groups hope that Congress will finalize the legislation next week, before lawmakers go on a 7-week summer recess. **http://scim.ag/GMfoodlabel**


Non-GM foods often bear labels; a new bill would require the same for GM foods.

Soil from a healthy ecosystem can help restore a degraded one.

## Soil 'booster shots' could turn barren lands green

Every good gardener knows that the right soil conditions are key to healthy plants. Now, ecologists have shown just how influential the right soil can be. Adding a thin layer of soil—and the organisms it contains—from healthy lands to denuded land can speed up recovery of the vegetation to years instead of decades, says E. R. Jasper Wubs, an ecologist at the Netherlands Institute of Ecology in Wageningen. The soil infusion can also influence what plants take hold in the once-barren land, he says. In the study, Wubs added soil from either a heathland, which has rolling hills of heather, or a grassland to plots of land where the topsoil was removed, along with seeds of 30 plant species. Six years later, plots receiving these boosters were richly carpeted with plants predominantly belonging to the plant community native to the original soil booster, Wubs and his colleagues reported this week in *Nature Plants*. They hope this approach can repair some once-fertile lands that are turning into desert. "It's a real-world demonstration that by altering soil, you can alter vegetation," says Noah Fierer, a microbial ecologist at the University of Colorado, Boulder. **http://scim.ag/_soilboost**

## Mega–health study launches

WASHINGTON, D.C. | President Barack Obama's ambitious 1-million-person personalized medicine study has begun to take shape. Last week, the White House announced $55 million in awards in 2016 for five recruitment centers and other components for the Precision Medicine Initiative Cohort Program, led by the National Institutes of Health. The program aims to recruit volunteers who are willing to share their health and genetic information over many years to help researchers develop individualized treatments. Four medical centers—Columbia University Medical Center; Northwestern University, Chicago, in Illinois; the University of Pittsburgh in Pennsylvania; and the University of Arizona, Tucson—will work with health care

PHOTOS: (TOP TO BOTTOM) R. NICHOLS/USDA NRCS; AP PHOTO/DAMIAN DOVARGANES/FILE

Chinstrap penguins on Zavodovski Island.

PHOTO: PETE BUCKTROUT, BRITISH ANTARCTIC SURVEY

## Penguin colony threatened by volcano's ash

Tiny, remote Zavodovski Island has more than 1 million chinstrap penguins (*Pygoscelis antarctica*)—and one volcano. The island is part of the South Sandwich archipelago, a group of volcanic islands in the southern Atlantic Ocean, about halfway between Argentina and Antarctica. Its residents make up the largest colony of chinstrap penguins in the world. But now, their numbers are threatened by their rumbling neighbor, Mount Curry, which has been erupting since March, as well as by a volcano dubbed Mount Sourabaya that is also erupting on nearby Bristol Island. Researchers became aware of the eruptions following a magnitude-7.2 earthquake last month. Fishing vessels in the area have snapped photos of Mount Curry, which is blowing smoke and ash eastward toward the penguins' habitat. The penguins are currently molting, which means they can't leave the island. Satellite images now reveal that nearly half the island has so far been covered in ash. "We don't know what impact the ash will have on the penguins," said geographer Peter Fretwell of the British Antarctic Survey in Cambridge, U.K., who has been involved in remapping the archipelago, in a statement. "If it has been heavy and widespread, it may have a serious effect on the population." Researchers plan to visit the archipelago later this year to take stock of the damage.

providers to recruit 150,000 volunteers each over 5 years starting in November. A fifth center at Scripps Research Institute in San Diego, California, aims to sign up 350,000 volunteers among the general public. **http://scim.ag/PMIlaunch**

## 1000 new cancer models

**BETHESDA, MARYLAND** | U.S. and European researchers are launching a 2- to 3-year pilot project that will produce 1000 new cutting-edge tumor tissue models for studying cancer. The Human Cancer Models Initiative will aim to give scientists an alternative to cancer cell lines, which often bear little resemblance to the tumor they came from. The project will draw on new insights into how to make the mixture of cells in a human tumor grow outside the body in a lab dish. Four groups are funding the effort: the U.S. National Cancer Institute, Cancer Research UK, the Wellcome Trust Sanger Institute in Hinxton, U.K., and the nonprofit Hubrecht Organoid Technology in Utrecht, the Netherlands. The first samples from the repository could be available this year. **http://scim.ag/cancermodels**

## Ag research nominee draws ire

**PARIS** | French scientists are sharply criticizing the nomination of a policy specialist to become the new president of the French National Institute for Agricultural Research (INRA). The critics say nominee Philippe Mauguin, who has served as chief of staff at the agriculture ministry since 2012, knows little about research and was offered the job as a political favor ahead of next year's general elections. Mauguin, who doesn't hold a Ph.D., was competing for the post against outgoing INRA President François Houllier, a former researcher. On 1 July, a collective called @INRAlerte released a 30-page statement to the press demanding that government officials withdraw Mauguin's application "because science is not a political reward." Several days later, the group launched an online petition in protest of the nomination that has gathered more than 2900 signatures. Agriculture Minister Stéphane Le Foll defended the nomination in the French Parliament last week, calling the accusations of conflict of interest "a lie." The French Parliament must still approve Mauguin's appointment to the 4-year INRA presidency, expected to begin on 26 July. If that happens, @INRAlerte says it would ask France's supreme court to block the appointment. **http://scim.ag/INRAnominee**

**NEWSMAKERS**

## Italian bird flu expert cleared

The legal travails of one of Italy's best-known scientists are over. Last week, a judge in Padua dismissed a series of criminal charges against veterinarian-scientist **Ilaria Capua**, now the head of the One Health

Center of Excellence for Research and Training at the University of Florida in Gainesville. The charges included spreading an infectious agent, criminal conspiracy aimed at corruption, handling stolen goods, and administration of drugs that endanger public health (*Science*, 5 September 2014, p. 1105). The judge decided that there is no case to answer. Best known for her work on avian influenza, Capua is a former division director at a governmental lab in Padua and a former member of the Italian Parliament. She had denied all charges and says she feels "relieved" but also "embittered" because the affair has harmed her credibility.

FINDINGS

## Cannibalism among Neandertals

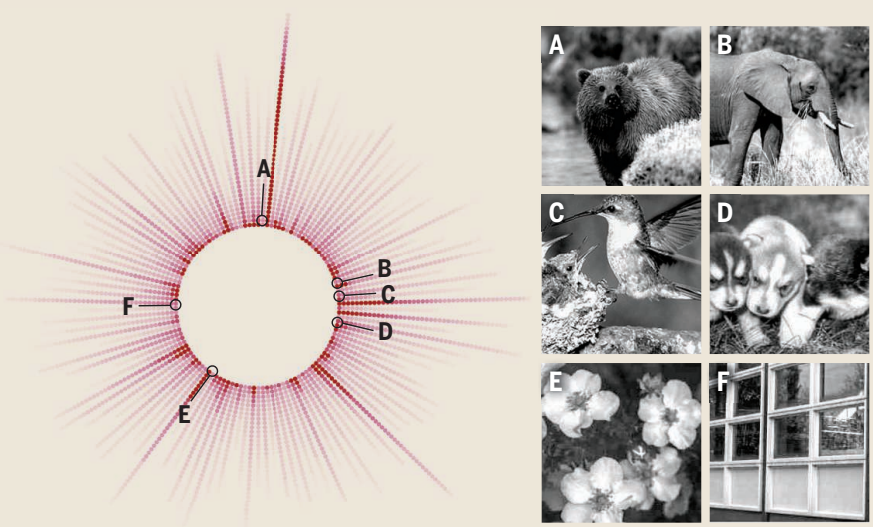
Neandertals ate each other—at least once in a while—according to a new analysis of bones unearthed in a Belgian cave.

A Neandertal boy, depicted in a museum diorama.

The remains were excavated near Goyet beginning in the 19th century and now sit in museums in Brussels. Many of the bones were covered in cut marks and dents caused by pounding, indicating that the meat and marrow had been removed, researchers reported last week in *Scientific Reports*. The scientists also spotted what appear to be bite marks running up and down finger bones. The marks were identical to those found on reindeer and horse bones also uncovered at the site, suggesting all three species were prepared and eaten. A few of the Neandertal bones showed additional wear and tear, suggesting they were later used to shape stone tools. The bones are between 40,500 and 45,500 years old, which is before *Homo sapiens* arrived in the region, so the only possible culprits are the Neandertals themselves. Although scientists knew that Neandertals had practiced cannibalism in Croatia, this is the first evidence of it in northern Europe. It's not clear whether Neandertals' cannibalism was a ritual practice, reserved for special occasions and imbued with special meaning—or whether they were just really, really hungry.

## How a mouse sees an elephant

The Allen Institute for Brain Science in Seattle, Washington, is best known for its brain atlases that reveal gene activity and neural circuitry in both mouse and human brains. But on 13 July, the institute released a more dynamic type of brain map that captures how more than 18,000 neurons in the mouse visual cortex respond to a variety of images, from simple black and white bars to a snippet of film. The responses were compiled from 35 individual mice. Understanding the brain requires more than just wiring diagrams of its circuits—researchers also need data (and lots of them) on how neurons communicate in real time, says neuroscientist Christof Koch, the institute's president and chief scientific officer. This foray into the brain's visual system is just the first installment of a long-term project to study how neural circuits perform the computations underlying perception and cognition, Koch says. From now on, the institute plans to release additional neural activity data four times a year, expanding to include different types of cells, targeted by the genes they express. Next up: additional brain regions involved in visual processing. And all of it will be freely available to other researchers.




Each line of dots represents neurons' responses to an image; darker dots are stronger responses.

PHOTOS: (TOP TO BOTTOM) © HEMIS/ALAMY STOCK PHOTO; ALLEN INSTITUTE

BY THE NUMBERS

**$173 million**

Amount that 890 U.S. cancer centers spent on advertising their clinical services in 2014; 59% of that spending was by the for-profit Cancer Treatment Centers of America (*JAMA Internal Medicine*).

**$750,000**

Total amount of a 3-year grant from the U.S. Office of Naval Research to a research team to help turn ordinary locusts—which have powerful olfactory systems—into bomb detectors.

**43%**

Fraction of people globally who do not know they are infected with HIV, found a report by the Joint United Nations Programme on HIV/AIDS. In the past 5 years, new HIV infections in adults have stopped declining.

DOPING

# *Cyclists' favorite drug falls flat in trial*

## EPO can break careers—but can it really make them?

*By* **Martin Enserink**,
*in Leiden, the Netherlands*

The grueling climb up Mont Ventoux in southern France is one of cycling's legendary ascents: a 22-kilometer path of pain, heroism, defeat, and death. One kilometer from the barren, windy summit is a memorial to Tom Simpson, the famous U.K. rider who collapsed and died there in 1967. An autopsy revealed that Simpson had taken amphetamines and alcohol to boost his performance, which may have contributed to sudden heart failure.

This year's edition of the Tour de France—which had a rendezvous with Mont Ventoux on Bastille Day, 2 days after *Science* went to press—has been remarkably free of doping scandals so far. But less than a month ago, in another race up the same slope, fully half the cyclists had a banned substance coursing through their veins. It wasn't a sports event, but a scientific study: the largest ever trial to test whether erythropoietin (EPO), a drug that has marred the Tour for decades, really enhances athletic performance as users claim.

The scientists who organized the study have yet to analyze the full results, but they have already shared a key outcome with the press: Riders on EPO, which stimulates the body to make extra oxygen-carrying red blood cells, were no faster than those who took a placebo. The finding flies in the face of previous, smaller studies and cycling lore, and some are incredulous. "Based on what we know so far about EPO, I would be very surprised if that's true," says Don Catlin, the CEO of Anti-Doping Research, a lab in Los Angeles, California.

But Catlin and other doping researchers applaud the team for subjecting EPO to a rigorous trial. Such studies are rare; critics have argued that the ever-growing list of banned substances in sports—and the international testing industry that enforces it—have no firm basis in science.

For the World Anti-Doping Agency (WADA) to ban a substance, it must meet two of three criteria: It enhances performance, poses a threat to athlete health, or "violates the spirit of sport"—a criterion that helped put recreational drugs like cannabis on the list. Proof isn't needed for any of the criteria, and evidence of performance enhancement is lacking in all but a handful of instances, such as the use of anabolic steroids in some strength-based sports.

Good studies are hard to do. Industry has no interest in footing the bill, and athletes would risk their careers if they participated. So a team led by Adam Cohen, the head of the Centre for Human Drug Research (CHDR) here, turned to amateurs. They enlisted 48 well-trained male amateur cyclists for a randomized, controlled, double-blind study—the gold standard in medicine—"as if we were testing a treatment for a disease called cycling slowly," says Jules Heuberger, a Ph.D. student at CHDR and a cycling aficionado. Cyclists from around the world offered to take part, he says. "EPO has this almost mythical status. People thought it would make them fly like a rocket."

Participants were injected with EPO or a placebo for 8 weeks, during which they completed seven endurance tests in the lab. The final challenge was a 130-kilometer ride ending in the race up Mont Ventoux on 19 June, a miserably cold day with an icy mistral blowing down the moonlike landscape. The winner, a lawyer, took an hour and 12 minutes to reach the summit. On average, riders on EPO were 38 seconds slower than the control group, a nonsignificant difference. And the EPO users hardly felt like they were running on rocket fuel; 62% guessed they were on a placebo.

But could EPO, the drug that allegedly helped Lance Armstrong rack up seven Tour victories, really be useless? It's conceivable, Cohen says. Riders may only remember winning with EPO and forget the losses, he notes. The drug could also be a strong placebo; after all, sports are a test of the mind as well as the body.

It's indisputable that the new red blood cells spawned by EPO increase the body's maximum oxygen uptake ($VO_2$max). But such an increase may not boost cycling performance, Cohen says. In long-distance

PHOTO: MICHIEL DE FOTOGRAAF 2015

Cyclists on erythropoietin were no faster than those on placebos in a race up Mont Ventoux last month.

races like the Tour, riders rarely perform at maximum uptake. Yet in a review of 13 previous EPO studies, the CHDR scientists found that all focused on $VO_2$max instead of performance. Some weren't controlled, only two had more than 20 participants, and none left the lab for a real-world race. "It's not what you would normally consider clear evidence," Cohen says.

Others are reserving judgment until the publication is in, but say they find it hard to believe EPO does nothing. There are biological reasons to suggest that increasing $VO_2$max has an effect, says Bengt Kayser, a doping expert at the University of Lausanne in Switzerland. Transfusions with a rider's own blood have been shown to boost $VO_2$max and performance, he says. And he notes that the body has EPO receptors in many places, including the brain, which might provide alternative mechanisms for the drug to have an effect.

Although testimonials from athletes themselves may not amount to hard evidence, they can't be dismissed, says Olivier de Hon, a scientist at the Anti-Doping Authority of the Netherlands in Capelle aan den IJssel, which advised on the study. Sports is full of fads that quickly fade—beet juice, rich in nitrates, is one recent example—but De Hon finds it hard to believe that cyclists would stick with a useless drug for decades. He says the scientists' time and money would have been better spent studying doping substances about which less is known, such as human growth hormone, insulin-like growth factor 1, and cocaine.

Cohen's group is now analyzing mountains of data on EPO. Among other things, they plan to look at its effects on blood clotting and muscle recovery. They will also send 300 urine samples from both groups to a WADA-accredited doping lab in Belgium to put EPO testing itself to the test. It's still possible that the EPO group performed better in the lab rides but, for some reason, failed to prevail on the mountain, Heuberger says. But he would not be surprised to find no difference at all.

Even if the negative finding holds up, EPO may stay on WADA's list. The agency tends to err on the side of prohibition, and politics plays into the process as well. Some have argued that it's fine to keep ineffective but potentially harmful substances on the list to deter athletes from using them. Cohen, on the other hand, believes delisting a drug will make it less attractive to athletes. If nothing else, he hopes last month's sprint up Mont Ventoux will show that a science-based approach to the issue is possible. ■

ANCIENT DNA

# First farmers' motley roots

## Unrelated groups adopted farming at about the same time in different parts of the Fertile Crescent

*By* **Ann Gibbons**

Ancient DNA has a way of uncovering complexity in seemingly simple stories of our past. Most famously, it has shown that modern humans didn't simply replace our archaic cousins as we spread across the world; we interbred with them along the way. Now, this method is adding nuance to the story of farming, long known to have originated in the Fertile Crescent of the Middle East.

According to three teams who used new techniques to gain glimpses of the nuclear DNA of the world's very first farmers, farming was adopted not by one group of people, but by genetically distinct groups scattered across the region. "It was not one early population that sowed the seeds of farming in western Asia, but several adjacent populations that all had the good fortune to live in the zone where potential plant and animal domesticates were to be found and exploited," says archaeologist Colin Renfrew of the University of Cambridge in the United Kingdom, who was not involved in the work.

The research—a paper published online in *Science* this week and two studies posted last month on the bioRxiv server—can't pin down whether agriculture spread quickly among diverse peoples or was independently invented more than once. But the diversity of the first farmers is "very surprising," says statistical geneticist Garrett Hellenthal of University College London, a co-author of the *Science* paper. "These early farmers who lived pretty close to each other were completely different."

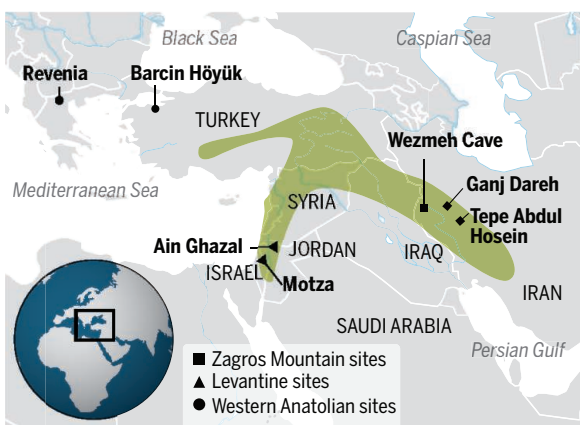
The earliest archaeological evidence for cultivating plants and herding animals dates back 10,000 to 12,000 years ago in the Fertile Crescent, which arcs from the Persian Gulf to Turkey and south to Egypt (see map, above). Excavations at Jericho in Jordan, Jarmo in Iraq, and Çatalhöyük in Turkey, for example, have found evidence of early grain farming and sheep and goat domestication in different areas at roughly the same time.

Geneticists have been trying to track whether one group of people—or just their ideas—spread farming early on. A single group did carry farming to Europe: DNA from ancient farmers in western Anatolia shows that they were the direct ancestors of Europe's first farmers, known as the Linear Pottery culture (*Science*, 11 October 2013, p. 181); present-day Sardinians share the most DNA with these ancient Anatolians.

But the trail of the first farmers went cold in the hot climate of the Middle East, which destroys DNA. Now, researchers are using new methods to prepare samples and extract them from the petrous bone of the ear, which is unusually rich in DNA. A team led by Joachim Burger of the Johannes Gutenberg University of Mainz in Germany and including Marjan Mashkour and Fereidoun Biglari of the National Museum of Iran in Tehran sequenced the complete genomes of four goat herders who lived in the Zagros Mountains of Iran. They include a 9000-year-old male from Wezmeh Cave, and three 10,000-year-old skeletons from a site called Tepe Abdul Hosein that are among the oldest remains of farmers in the world. Tests of isotopes in the teeth of all four farmers confirm they had diets rich in grains, a signature of an agricultural diet.

### Who farmed first

DNA from the world's oldest farmers in the Fertile Crescent (green) shows that they were a diverse lot from the beginning, with distinct genetic groups in Iran, Israel/Jordan, and western Anatolia/Greece.




CREDITS: (GRAPHIC) V. ALTOUNIAN/*SCIENCE*; (DATA) J. BURGER

By using a new method that looks at patterns of inheritance of chunks of DNA, Hellenthal found that the early Zagros Mountain farmers have left a genetic legacy in Pakistanis, Afghans, and others, particularly in Zoroastrians in Iran.

But the ancient Iranian DNA was dramatically different from that of the western Anatolian farmers. The two groups of farmers, who lived about 2000 kilometers and 2000 years apart, must have descended from completely different groups of hunter-gatherers who separated 46,000 to 77,000 years ago, Burger says.

A similar genetic disjunction appears in a study led by Harvard University's David Reich and posted on bioRxiv. This study analyzed ancient DNA from 44 Middle Easterners who lived 14,000 to 3400 years ago, including Natufian hunter-gatherers in Israel, Zagros farmers, and Bronze Age pastoralists in the Eurasian steppe, and compared it with that of 2864 living and ancient people from around the world. By sequencing 1.2 million nucleotides from across each genome, the team found that early farmers of Israel and Jordan (known as the Levant) were genetically distinct from those in the Zagros Mountains, and that both populations were distinct from the western Anatolians who later spread their genes throughout Europe.

The third study, also published on bioRxiv, reported the same stark differences. That study analyzed the complete genome of a 10,000-year-old woman from Ganj Dareh, a site in the Zagros Mountains with the world's oldest evidence of goat herding.

Burger and Reich also each used their data to peer even further back in time, to the ancestors of the Zagros Mountain farmers. They found that the Zagros people descend from a group of basal Eurasians who separated from the ancestors of all other people outside of Africa 50,000 to 60,000 years ago—before other non-Africans interbred with Neandertals. So the Zagros Mountain farmers had less Neandertal DNA than the western Anatolian farmers, whose ancestors must have branched off later.

The descendants of these early farmers went separate ways. Whereas the western Anatolians later migrated to Europe, Reich's team proposes that the ancient farmers of the Levant migrated to East Africa, where living people carry some of their distinct DNA, and the Zagros Mountain farmers spread north into the Eurasian steppe and east into South Asia.

Did these early people learn farming from each other, or was it invented more than once? Here, opinion differs. Archaeologists have noted that early farmers in different regions used different tools and grains, supporting the idea of multiple origins, says archaeologist Roger Matthews of the University of Reading in the United Kingdom. "The genetic and archaeological evidence suggest at least two separate pathways to agriculture, at distant ends of the Fertile Crescent, eventually merging into a unified package that then spreads outwards," he says.

But these groups traded obsidian, suggesting to Renfrew and Harvard archaeologist Ofer Bar-Yosef that seeds and farming knowledge could have been shared, too. Because new kinds of food preparation tools turn up first in the Levant, Bar-Yosef thinks farming sprouted here: "Zagros foothills people adopted agriculture from the Levant."

Burger suggests that farming was such an advantage that it spread both as an idea and by migration of people. "Initially, agriculture was an idea that spread," he proposes. "Then, when it reaches the borders of Europe, it becomes people spreading farming. We have an extremely complex agricultural revolution that was created by people who were extremely diverse." ■

Zoroastrians, who practice Iran's ancient religion, still carry DNA from the earliest farmers in the Zagros Mountains.

PHOTO: VAHID SALEMI/ASSOCIATED PRESS

NEUROSCIENCE

# Brain scans are prone to false positives, study says

## Common software settings may have skewed the statistics for thousands of studies

*By* **Greg Miller**

Two neuroscientists, functional magnetic resonance imaging (fMRI) has delivered both unmatched insights and occasional doses of embarrassment. In the 1990s, fMRI revolutionized the field by providing a way to study brain activity in human subjects. More recently, however, a handful of studies have pointed to flaws in how fMRI researchers analyze their data and in some of the assumptions underlying the technique, which infers brain activity from changes in blood flow. Media coverage of both the field's more sensational "this is your brain on politics"–type findings and its woes (such as the "voodoo correlations" fracas a few years ago) haven't helped its credibility.

Now, the field is buzzing about an analysis published online 28 June in the *Proceedings of the National Academy of Sciences* (*PNAS*). Anders Eklund, an electrical engineer at Linköping University in Sweden, and colleagues examined statistical methods in three software packages commonly used to analyze fMRI data. They found that certain common settings in the software gave rise to a false positive result up to 70% of the time. In the context of a typical fMRI experiment, that could lead researchers to wrongly conclude that activity in a certain area of the brain plays a role in a cognitive function such as perception or memory.

Some other neuroscientists note that these pitfalls in fMRI interpretation have been known for decades, and savvy researchers know how to avoid them. But Geoffrey Aguirre, a neuroscientist and neurologist at the University of Pennsylvania, says the new study has demonstrated the problem "more clearly than anyone had done before, using a larger data set than has been used before." When Eklund and colleagues posted their

A new study has renewed the debate over the use of statistics in functional magnetic resonance imaging research.

PHOTO: JAMES KING-HOLMES/SCIENCE PHOTO LIBRARY

paper on arXiv.org last November, they suggested that their findings raised questions about all 40,000 fMRI studies published in the last 2 decades, but they've reduced the number after a closer look at the literature. In a blog post last week, Eklund's co-author Thomas Nichols, a statistician at the University of Warwick in Coventry, U.K., estimated that about 3500 studies used the problematic software settings—still an alarming number.

The problem arises from a fundamental challenge in fMRI research. In a typical experiment, subjects do a task inside the scanner, something to engage their memory, attention, social skills, or whatever interests the researchers, while the scanner monitors their brain activity. Then, the researchers have to determine whether the patches of activity picked up by the scanner are really related to the experiment or just a random fluctuation.

One common solution is a statistical method called cluster-wise inference, which considers both the strength of activity at spots throughout the brain as well as the size of those spots. Cluster-wise inference is built into the three popular software packages analyzed in the study—statistical parametric mapping (SPM), the fMRI of the brain Software Library (FSL), and analysis of functional neuroimages (AFNI). All enable researchers to set a few parameters to adjust how the analysis is done and how stringent it will be. The problem, Eklund and colleagues say, is that when one of these parameters—something called the cluster-defining threshold (CDT)—is set too low, the analysis is more likely to result in a false positive.

In the new study, the researchers drew on several public databases that contain fMRI data collected while subjects were resting in the scanner, not engaged in any particular task. The researchers analyzed those data as if they were running a typical fMRI experiment, looking for regions of brain activation related to a task. The team simulated nearly 3 million fMRI experiments. Based on the statistical threshold they'd set, they expected to get a false positive result (that is, a positive hit for task-related activity even though there was no task) 5% of the time. Instead, depending on the software and the settings, up to 70% of the results were positive. They also identified a bug in the AFNI software package that had existed for 15 years and may have contributed to false positives. The team alerted the developers last year, and the bug has been fixed.

> ***"You'd hope that when we build a whole [scientific] field that the fundamental tools would have been validated ..."***
>
> **Russell Poldrack**, Stanford University

The pitfalls identified in the *PNAS* paper are nothing new, says Karl Friston, who directs the Wellcome Trust Centre for Neuroimaging at University College London and pioneered the development of statistical methods and software for analyzing fMRI data. Indeed, Friston and colleagues first identified these problems in a 1994 paper. In a technical response published on arXiv.org last month, Friston and Guillaume Flandin, also at the center, argue that the trouble can be avoided by choosing more conservative statistical thresholds. "Scientifically, I have nothing to add," Friston wrote in an email. But the work has continued to generate discussion elsewhere, from an initial frenzy on Twitter when Eklund and colleagues published their preprint to last month's Human Brain Mapping meeting in Geneva, Switzerland.

Much of the software used in fMRI research hasn't been validated with actual data, as Eklund and colleagues have done, says Russell Poldrack, a neuroscientist at Stanford University in Palo Alto, California. "You'd hope that when we build a whole [scientific] field that the fundamental tools would have been validated with real data, not just theory and simulation," he says. "It took 20 years to happen."

In this case, it seems that both researchers and the software developers deserve some of the blame. "[CDT] is a parameter that the user must set, but they are presented with a default that they can accept, and most do," says Nichols, who is on the development team for two of the software packages, SPM and FSL. Both packages are being modified to discourage users from picking the problematic settings.

What this means for the last 2 decades of fMRI research isn't entirely clear. "Just because the statistical significance of a particular finding is overestimated doesn't automatically mean the scientific findings of the paper are wrong," Nichols says. In other words, even if 3500 papers used the problematic methods, the number of invalid scientific findings is certainly less.

It will largely be up to the original labs to reanalyze the work—if they choose to. "We would hope that researchers would be interested to know whether their previous claims stand, but realistically there is very little incentive (and lots of disincentives) to show that one's previous results are wrong," Poldrack says.

The new work is yet another reason fMRI researchers should share their data more freely than they do, adds Poldrack, who in 2010 established the OpenfMRI repository, one of the sources for the data used in the new study. Data sharing not only enables researchers to try to replicate one another's findings, but also makes it possible to re-evaluate old studies when a methodological flaw comes to light. "Without widespread data sharing," Poldrack says, "it's basically impossible to know whether any particular finding is robust to these issues." ■

*Greg Miller is a science and technology journalist based in Portland, Oregon.*

ARCHAEOLOGY

# A time capsule from Bronze Age Britain

## Charred river houses offer an extraordinary view of everyday life 3000 years ago

*By* **Erik Stokstad**

Most likely, a raiding party torched the village. Flames raced through the wooden roundhouses so quickly that the inhabitants fled without their belongings. The scorched remains, perched on stilts, eventually collapsed into the river below. Buried in silt, the debris remained intact for 3000 years, preserving a remarkable record of ordinary life in the Late Bronze Age.

Now, it is coming to light. This month, archaeologists are wrapping up the £1.4 million excavation near Peterborough, U.K. Next, they will begin analyzing the many artifacts—textiles and wooden objects preserved by charring and burial in the sediments, along with pottery and metal tools such as chisels and tweezers. "You've got the closest thing to a Bronze Age Pompeii," says Benjamin Roberts, an archaeologist at Durham University in the United Kingdom who is not involved in the excavation.

"Britain's Pompeii" has yielded a treasure trove of intact pottery and other household items.

In the United Kingdom, the Bronze Age began between 4500 and 4000 years ago with the expansion of large monuments like Stonehenge. By the Late Bronze Age, farming was a focus of innovation. People developed extensive systems of fields and dug watering holes for livestock. Life wasn't entirely bucolic; bronze swords and spears at many sites testify to a threat of violence. But domestic details are scarce. Little is left of Bronze Age houses except postholes and hearths.

The first clues that Must Farm, about 120 kilometers north of London, might reveal more were oak posts poking from beds of clay at a brick quarry. Tree ring analysis dated the posts to 1290 to 1250 B.C.E., and excitement grew when preliminary digs unearthed intact pottery and rare Bronze Age textiles. Wooden boats hinted at the vast wetlands that once covered the region.

The ancient wood was at risk in the brick quarry, so a major excavation began last October. Since then, the Cambridge Archaeological Unit (CAU) in the United Kingdom, a contract firm, has found at least five roundhouses that likely date to 900 B.C.E. (The older posts appear to have belonged to a causeway across the river that had fallen into disrepair by the time the houses were built several centuries later.) These stilt houses may have been built over the river to foster trade or for defense.

Each roundhouse had a rich inventory of domestic items. "I've been an archaeologist for 30 years and I've never seen anything this spectacular or exciting," says Stephen Trow, research director of Historic England at Fort Cumberland, a government agency that is helping fund the excavation with the brick company.

The arrangement of artifacts could indicate how various sections of the houses were used. All the structures contain a similar set of ceramics—ranging from tiny cups to fine bowls and coarse storage jars—which may reveal areas for cooking, eating, or storing food. Butchered lambs, cut in half along the spine, were found inside one quarter of a house, whereas the remains of wild pigs and deer had been dumped in the river, possibly reflecting a taboo against butchering wild animals indoors, says Mark Knight, CAU's excavation manager.

Many vessels and pots contain charred grains of barley, wheat, and residues of cooked food. One bowl even sank with a spoon still lodged in the burnt crust of a stew. Chemical analyses "will almost allow us to get Bronze Age recipes," says Alexander Gibson of the University of Bradford in the United Kingdom, who is not participating in the research.

Must Farm's textiles are an especially precious find. The site has provided the largest collection of Bronze Age fibers and fabrics in the United Kingdom, says Susanna Harris at the University of Glasgow in the United Kingdom, who is collaborating with the team. Artifacts include balls of thread, hanks of yarn, and fine linen. Thread counts are as high as 30 per centimeter, comparable to the best cloth known in Europe at the time. "I counted them several times, thinking 'This can't be right,'" Harris says. So far she has not seen clues, such as cuffs, that would reveal whether the fabrics were part of clothing or served other functions.

The houses were relatively new at the time of their destruction. A wooden palisade had also recently been built, as indicated by fresh wood chips, green timber, and a lack of insect damage. Knight thinks the barrier was intended to protect the houses. "I'm absolutely convinced it's saying, 'Keep out. We'll defend ourselves.'"

Karl Harrison, an archaeologist at Cranfield University in the United Kingdom, is studying the scorch marks and other fire damage. He wonders whether the fire started inside a house—perhaps from a cook fire—or outside, which might point to arson, as many researchers suspect. Either way, the blaze appears to have spread quickly. "It was rapid, smoke-filled, and incredibly destructive," he says. "You'd have a couple of minutes to scrabble around."

The inhabitants apparently escaped, as no human skeletons have been found in the debris. They never returned to rebuild, which converted their misfortune into a stroke of luck for archaeologists today. ■

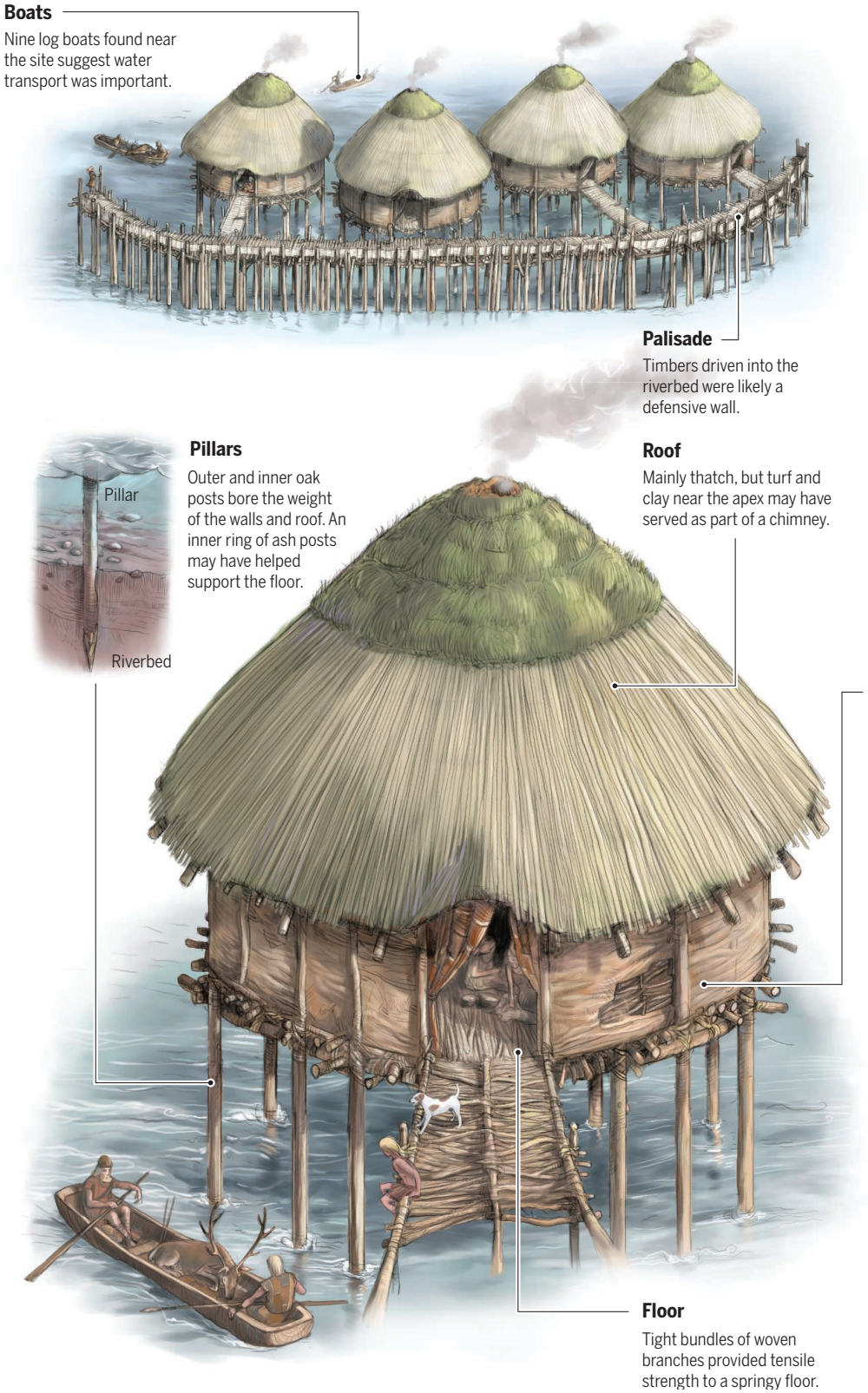
**ONLINE**

View an interactive version of the graphic at http://scim.ag/29ztLAk.

PHOTO: DAVE WEBB/CAMBRIDGE ARCHAEOLOGICAL UNIT

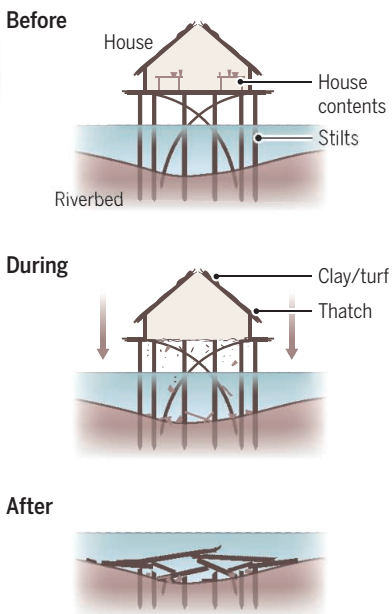
## A river ran through it

Built over a wide river with access to the North Sea and inland farms, these Bronze Age houses were well-positioned perches for people to trade grain, meat, and metal tools. Yet not long after they were built, the houses burned down and collapsed into the water, which preserved their contents. Perhaps it was arson by a jealous chiefdom or cattle rustlers—archaeologists may never identify the culprit, but they are learning much about daily life here in the British Isles 3000 years ago.




### A fiery end

Fire rapidly engulfed the houses. The floors soon gave way, dropping the contents into the water. Eventually the walls and roofs collapsed.




### The dig

Archaeologists spent 10 months removing objects and documenting the jumble of fallen timbers.

FEATURES

# THE AVENGER

## Afflicted by a rare disease, a young doctor fights back

*By* **Jennifer Couzin-Frankel**, *in Philadelphia, Pennsylvania*

After the third time he nearly died but before the fourth, David Fajgenbaum embraced a new motto: Think it, do it. "I got out of the hospital with this profound sense of, you need to make the most of every second," he says. A former college football player with close-cropped dark hair and a firm handshake, Fajgenbaum, 31, is the picture of youthful vigor now. But that belies a frightening reality. Tomorrow the symptoms with which he's all too familiar could return, sending him to the intensive care unit (ICU) with every organ failing.

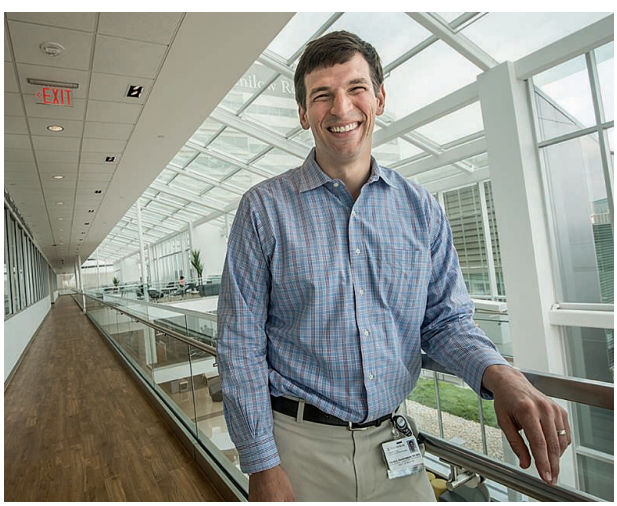
Fajgenbaum was in his third year of medical school 6 years ago, on an obstetrics-gynecology (OB-GYN) rotation, when he was first hit by night sweats, fatigue, and weight loss. "I didn't know what it was, but I knew something was terribly wrong," he says. As his health deteriorated, he gritted his teeth and kept on delivering babies. After sitting through his OB-GYN exam, he stumbled a few blocks to the emergency room at the Hospital of the University of Pennsylvania (UPenn).

"They ran some bloodwork and they said, 'Dave, your liver, your kidneys, and your bone marrow are not working, what's going on?'" Fajgenbaum was hospitalized straight away. A retinal hemorrhage causing temporary blindness, along with fading organ function, landed him in the ICU. There he remained, without a diagnosis, for nearly 7 weeks. Eventually Fajgenbaum recovered, and his baffled doctors "kept saying, 'Let's just hope it doesn't come back.' We had no idea what it was."

Fajgenbaum spent his first weeks out of the hospital combing through thousands of pages of medical records stretching back to childhood, hunting for something, anything, that doctors might have missed. As his medical saga unfolded, he learned that his potential killer was a rare and vicious immune disorder—Castleman disease, which strikes about 5000 people in the United States each year—and knowledge about it was in depressingly short supply.

Fajgenbaum abandoned plans to become an oncologist, skipped medical residency, and enrolled in business school instead—building a powerhouse network of hundreds of physicians, researchers, and drug company employees around the world to help him decipher Castleman. He co-authored papers with his doctor, wrote a case study about himself, proposed a new model of the disease, and currently coordinates a dozen Castleman studies from his small office at UPenn, where he is an assistant professor. His bare chest, pockmarked with blood moles that are a harbinger of a Castleman relapse, appeared on the cover of *JAMA Dermatology* in early 2013. It's not exactly Mr. February, but Fajgenbaum is proud.

He acknowledges that he might die before completing his quest to understand and cure his affliction. "If this does come back and I don't survive that episode, [I] would be sad … that I won't be with my family and friends," he says evenly. But "I'm literally doing everything humanly possible" to save myself and many others. And somewhat to his surprise, he's crazy about his work. "I'm enjoying this battle," he says. "I'm enjoying the chase after the unknown."




From his base at the University of Pennsylvania, David Fajgenbaum is trying to change the prospects for a dire disease—his own.

PHOTO: © APRIL SAUL

**A BOSTON PATHOLOGIST** first described Castleman disease in 1954. Benjamin Castleman was mystified by a 40-year-old executive with fever and weakness who had an unusual mass in his chest. Doctors initially suspected cancer or another disorder, but the mass, when surgically removed, turned out to be something else altogether—what would later become known as unicentric, or localized, Castleman disease.

The disease was considered a sort of "prelymphoma" with autoimmune features, in which cells in lymph nodes proliferate and attack various tissues. Over time, doctors pinpointed an immune-system messenger called interleukin-6 (IL-6) as key. They came to believe that Castleman flares when lymph nodes secrete excessive IL-6, setting off a massive inflammatory reaction, either locally or throughout the body.

Fajgenbaum was diagnosed with the disorder's most dangerous subtype, idiopathic multicentric Castleman disease (MCD), during his second bout of flulike symptoms and organ failure, which hit while he was still recovering from the first at his family's home in Raleigh. A lymph node biopsy revealed the typical Castleman mix of atrophy and enlargement in different parts of the node, along with blood vessel overgrowth. Physicians turned to their usual rescue for critically ill Castleman patients: aggressive chemotherapy, intended to wipe

David Fajgenbaum during treatment in 2012. It took a mix of seven chemotherapy drugs to save him.

PHOTO: © SIMON GRIFFITHS

out the immune cells that were attacking healthy tissue.

Despite treatment, Fajgenbaum languished, and doctors told his Catholic family that he would not survive. A priest was called to offer last rites. It was November 2010, and Fajgenbaum was 25 years old.

Two days later, his body began to right itself. In an effort to safeguard his health, he decided to travel west to seek counsel from a world expert in Castleman disease, Frits van Rhee at the University of Arkansas for Medical Sciences in Little Rock. During that visit his symptoms returned. This time he was hit with seven chemotherapy drugs. Another brush with death was followed by slow recovery.

Despite spending 4.5 of the previous 6 months hospitalized, Fajgenbaum was feeling more upbeat. "I just trusted that the medical system was going to figure this out," he says. The research papers he'd skimmed were reassuring. "People were speaking very dogmatically about how the disease works. I just assumed that if this is in the medical literature, this is the case."

His optimism turned out to be misplaced. Fifteen months later, the familiar symptoms returned, and Fajgenbaum settled into a Little Rock hospital room for another extended stay and more chemotherapy. And now he began taking matters into his own hands, pressing Van Rhee for answers. Which immune cells were behind this? Which pathway? Why wasn't treatment working? What was the contingency plan?

Van Rhee was unable to comment on this story before *Science* went to press. But Fajgenbaum recalls that his doctor's answer was not comforting. "He said, 'Dave, we've tried everything on you and you're not responding to anything. What do you think we should do?'" At that point, "it hit me like a ton of bricks," Fajgenbaum says. "I really realized how far away we were from the goal line."

With his fiancée and two older sisters camped out next to him, Fajgenbaum made a promise: He would spend the rest of his life, however much remained, "trying to take this thing down." The same seven-drug chemo cocktail saved him, and on 2 June 2012 he left the hospital. "I don't think I've stopped sprinting since," he says.

**AT FIRST,** Fajgenbaum had to figure out in which direction to run. Guidance came from another young man, Josh Sommer, now 28 years old, who was diagnosed as a freshman at Duke University with chordoma, a rare bone cancer that often appears in the skull and spine. Sommer later dropped out of college to form the Chordoma Foundation in Durham, North Carolina, which he's headed ever since.

The two connected through a mutual friend in 2012 and met at a Raleigh coffee shop. Sommer had a message for Fajgenbaum: Every rare disease needs a quarterback, someone to marshal the team, harness the resources, and lay out a game plan. "People have to put their faith in this person," Sommer says, to drive a sometimes rickety research train toward a clear destination. "I said, 'Look … this disease needs you and you can make a huge impact,'" Sommer recalls. To garner respect among scientists, Fajgengaum would have to become literate in the disease—"a peer," Sommer told him. The two talked for hours.

Meanwhile, Fajgenbaum's initial hopefulness about his prognosis continued to fade. One trigger was his senior project for medical school, which began with a literature review of Castleman publications.

There was one mystery about his own case: Upon his most recent relapse, he was already taking a drug that blocks IL-6, which was then experimental and has since become the only approved treatment for MCD. But if IL-6 causes the disease to flare, and Fajgenbaum had none, how could Castleman come back? The papers deepened his confusion. Fajgenbaum came to believe that the accepted model—Castleman is caused by tumors secreting IL-6, which sends the immune system into overdrive—was unsupported by actual data.

One day, as he continued his senior project detective work, Fajgenbaum noticed something odd as he pored over images of lymph nodes from patients with autoimmune diseases, like lupus and rheumatoid arthritis. "They had almost identical lymph node features" to Castleman, he says. "It was so peculiar." In Castleman, the lymph node abnormalities are considered drivers of the disease; in lupus, they are a reaction or side effect. Could it be, he wondered, that as in lupus, enlarged lymph nodes are an outcome of the disease, rather than its cause?

Fajgenbaum needed the medical community's help. He scoured the PubMed database to find every researcher who had published on Castleman, and wrote more than 400 personal emails, inviting them to a meeting he was organizing at that December's American Society of Hematology conference in Atlanta. To his surprise, many seemed unaware of each other and their work.

One of those who said yes was Jean-François Rossi, a hematologist at Saint-Eloi Hospital in Montpellier, France. Castleman scientists "are from different origins," Rossi explains, spanning hematology, multiple myeloma, lymphomas, and other specialties. Rossi was curious to learn more and eager to meet with other experts.

In Atlanta, 27 Castleman experts gathered in a conference room in the convention center, with another five dialing in over the phone. Fajgenbaum, an unknown medical student 6 months from graduation, ran the event. There were disagreements about disease terminology, about which cells were worth studying. "It became so clear there was no consensus," says Fajgenbaum, who delicately calls the gathering "revealing."

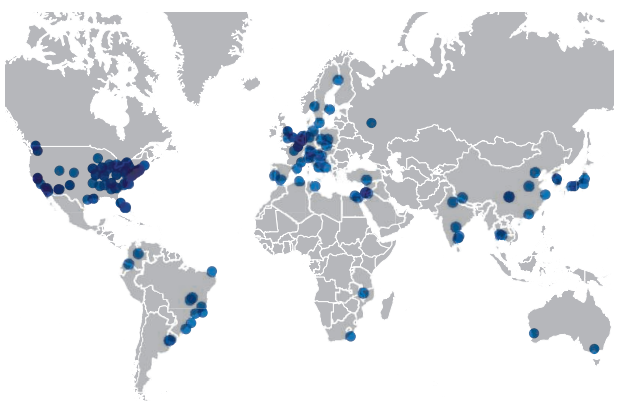
The challenges, Fajgenbaum came to believe, weren't rooted in science so much as business. "There was no overarching strategy," he says, no 5-year road map of the kind corporations embrace. He concluded that to tackle Castleman he needed to pursue an MBA, and successfully applied to one of the world's best programs, at the nearby Wharton School at UPenn.

**"I REMEMBER WHEN** he came to interview," says June Kinney, who teaches health care management at Wharton. "He looked very vigorous and healthy. … How could he look so healthy and have this shadow hanging over him?"

Kinney was impressed not only by Fajgenbaum's current plans but by a tragedy-fueled achievement in his past. When Fajgenbaum was 19, his mother had died of brain cancer. Devastated and isolated,

## Networking the globe

Reaching across national borders, the Castleman Disease Collaborative Network is linking up doctors and researchers, including some who haven't worked in the disease before.



DATA: COURTESY OF DR. DAVID FAJGENBAUM

he launched an organization for grieving college students called AMF—which then stood for Ailing Mothers and Fathers, and also his mother's initials, Anne Marie Fajgenbaum. Since renamed Actively Moving Forward, AMF now has 55 chapters at colleges around the country. Fajgenbaum continued to head the sprawling organization through medical school. This year, he coauthored a book, *We Get It: Voices of Grieving College Students and Young Adults*, a collection of 33 personal narratives.

Launching and running the nonprofit "was the ultimate training ground" for overhauling the Castleman world, he says. Kinney agreed. "I don't get very many people that come in and have that kind of accomplishment at that stage," she says. "Obviously there's a lot of innate talent there in terms of motivating and inspiring people to build an organization."

At Wharton, Fajgenbaum shoehorned his class commitments around his sequel to AMF: the Castleman Disease Collaborative Network (CDCN). He also drew on class lessons on, for instance, negotiation strategies, the economics of drug development, decision-making, and management.

During his first semester in business school, Fajgenbaum, his physician Van Rhee, and a medical school friend, Christopher Nabel, wrote a paper for *Blood* that proposed the new model for his form of Castleman disease. The trio suggested that the lymph node effects were secondary, and that the disease was driven by some sort of systemic inflammatory disease mechanism, like certain gene mutations, or autoantibodies, or possibly a virus. (One virus, called HHV-8, is known to be behind a subset of MCD cases, but Fajgenbaum's form is idiopathic.)

Fajgenbaum was keenly aware that his own treatment matched the traditional model, in which IL-6–secreting tumors are the disease driver. He was still taking the IL-6 blocker that had failed him before, as well as some maintenance chemotherapy. As he readied the *Blood* paper for publication, the disease returned, and he came perilously close to death for a fifth time.

The experience prodded him to overhaul his own treatment. Based on studies of his own blood and his rethinking of the disease, he wondered whether a drug suppressing activated T cells might work. His doctors agreed, and he became, to his knowledge, the only Castleman patient taking a particular immunosuppressant approved for other conditions.

In a sense, each relapse had brought an epiphany—he needed Castleman experts, he needed to take on this disease himself. This time, "I was thinking to myself, I've spent every moment of every day, and it just wasn't enough. … Maybe it's that we need to make this bigger than Castleman disease, engage people outside" the field. He left the hospital 4 weeks later considering how to recruit more bright minds to the fight.

**"I'M A COG** in their machine," says W. Ian Lipkin, a renowned microbe hunter who runs a 50-person lab at Columbia University. Lipkin was familiar with the viral form of Castleman disease. But he had no idea there were Castleman patients, like Fajgenbaum, without any obvious sign of that virus.

Lipkin fits neatly into Fajgenbaum's grand design: Identify the most pressing Castleman projects, then recruit the best person for each job. Fajgenbaum connected hundreds of researchers through an online portal, crowdsourced for a list of possible Castleman studies, then prioritized them—a strategy inspired by a Wharton exercise in which all 800 first-year business students had to collectively come up with five approaches to combat climate change.

> ***"I got out of the hospital with this profound sense of, you need to make the most of every second."***
>
> **David Fajgenbaum**, University of Pennsylvania

The final tally included about 20 studies. One involved sifting through tissue samples from patients and looking for viruses, bacteria, or other nonhuman DNA or RNA that might trigger the disease. "Once we knew we wanted to do pathogen discovery, the first thing was an inventory" of every expert in the world, Fajgenbaum says. He and Nabel tapped immunologists, rheumatologists, and deans—"people who have a really good pulse." Lipkin's name rose to the top.

Lipkin's lab hums along with more than $4 million a year in federal funding. So the $57,000 CDCN could offer him was unlikely to be much of an incentive. But the young doctors had a connection. Chris Nabel's father is Gary Nabel, an old friend of Lipkin's and former head of the Vaccine Research Center at the National Institutes of Health.

"Gary called me in June of 2014 and asked whether or not I could help his son and some of his friends," Lipkin says. "After I spoke with these kids, I got involved because they're earnest and committed."

Lipkin asked for 25 frozen lymph nodes from 25 different patients, which seemed an easy request because everyone with Castleman gets a lymph node removed. Then Fajgenbaum and Nabel discovered that in more than 95% of cases, the lymph nodes are embedded in paraffin and kept at room temperature—not stashed in a freezer as Lipkin needed for analysis.

The global network of researchers, which then numbered about 350, came through. Two physicians in Japan and another in Norway had some samples to offer. Eventually CDCN pulled together 34 samples for Lipkin. There followed material transfer agreements and more paperwork. It took 14 months from the time the lymph node hunt began until all the samples arrived at Columbia, for a study slated to last 3 months that should be complete any day now.

"I think David is doing this in a very modern way," says Alexander Fosså, a medical oncologist at Oslo University Hospital and the Norwegian who, to Fajgenbaum's delight, had frozen samples. For Fosså, as for many other physicians, Castleman is "like a small niche" that he's pursued since residency, when he struggled to treat his first patient. "When I do this as a hobby, I need to have some positive experience," Fosså explains. "That's what [Fajgenbaum] guarantees me—he's fun talking to, he's fun emailing with; that makes me smile."

Fosså holds that IL-6 is key to the disease, as evidenced by a clinical trial he was involved in, which led to the approval of an antibody drug, siltuximab, for Castleman in 2014. But he agrees it's not the underlying cause. Studies to identify other mechanisms are moving forward, including genome sequencing of about 10 Castleman patients and their parents, and comparing levels of proteins during flares and remissions.

In February, CDCN, UPenn, and a drug company inked an agreement to set up a patient registry; the cost and the company's identity are still under wraps. The goal is to better understand the trajectory of Castleman, examine which treatments patients are taking, and identify those that might work.

High on that list is Fajgenbaum's own experimental treatment. Fajgenbaum has been on his immunosuppressant for 2.5 years now and remains healthy. The patient in him wants to "scream from the mountaintop," telling everyone to try this treatment. The doctor in him is cautious and declines to even name it, for fear his good health is due to chance, not biology. If he remains relapse-free at 3 years—for another 6 months or so—then he plans to submit a case study describing his experience.

"I need to be absolutely prudent," Fajgenbaum says. He's not just another Castleman patient. ■





ECOLOGY

# *Butterfly communities under threat*

## Butterfly populations are declining worldwide as a result of habitat loss and degradation

*By* **Jeremy A. Thomas**

Butterflies are better documented and monitored worldwide than any other nonpest taxon of insects (*1*). In the United Kingdom alone, volunteer recorders have sampled more than 750,000 km of repeat transects since 1976, equivalent to walking to the Moon and back counting butterflies (*2*). Such programs are revealing regional extinctions and population declines that began before 1900 (*3, 4*). In a recent study, Habel *et al.* report a similar story based on inventories of butterflies and burnet moths since 1840 in a protected area in Bavaria, Germany (*5*). The results reveal severe species losses: Scarce, specialized butterflies have largely disappeared, leaving ecosystems dominated by common generalist ones. Similar trends are seen across Europe (*6*) and beyond, with protected areas failing to conserve many species for which they were once famed.

The butterfly losses are severe. On the Bavarian reserve, Habel *et al.* found that 71 species survive compared with 117 in 1840 (*5*). In the Netherlands and the English county of Suffolk, 24 and 42%, respectively, of resident breeding species became extinct during the late 19th and 20th centuries—an order of magnitude more than the population declines and regional and national extinctions of native vascular plant and breeding bird species (*4*). Baseline data for insects are sparse outside Europe. Nevertheless, there is evidence for similar declines in North America, Japan, and hotspots of butterfly endemism such as Brazil, South Africa, and Australia, not

*Department of Zoology, University of Oxford, Oxford OX1 3PS, UK. Email: jeremy.thomas@zoo.ox.ac.uk*

PHOTO: DAVID SIMCOX/HABITAT DESIGNS LTD

Following extinction in the United Kingdom in 1979, the globally threatened large blue (*Maculinea arion*, see photo) was successfully reintroduced using a similar genotype from Sweden. It has now spread to about 35 protected areas (*16*).

least among the iconic birdwings of the Indo-Australian tropics (*7–11*).

Butterflies constitute just 2% of the world's 950,000 described insect species, but it is increasingly evident that their rates of loss are matched—and even exceeded—by other groups, including bumblebees, dragonflies, moths, and ladybirds, whose respective social, aquatic, nocturnal, and predatory lifestyles make them particularly susceptible to environmental changes caused by humans (*4, 12, 13*). This matters because bees and moths are essential pollinators of numerous plants, including crops, while ladybirds are valued predators of insect pests. Butterflies, by comparison, contribute little to landscape functioning or ecosystem services. Their value for humans is largely aesthetic and as indicators of diversity. Yet, even here there are exceptions, such as the mountain pride (*Aeropetes tulbaghia*), the sole known pollinator of about 20 endemic plant species of southern Africa (*10*).

## LOSS OF BREEDING AREAS

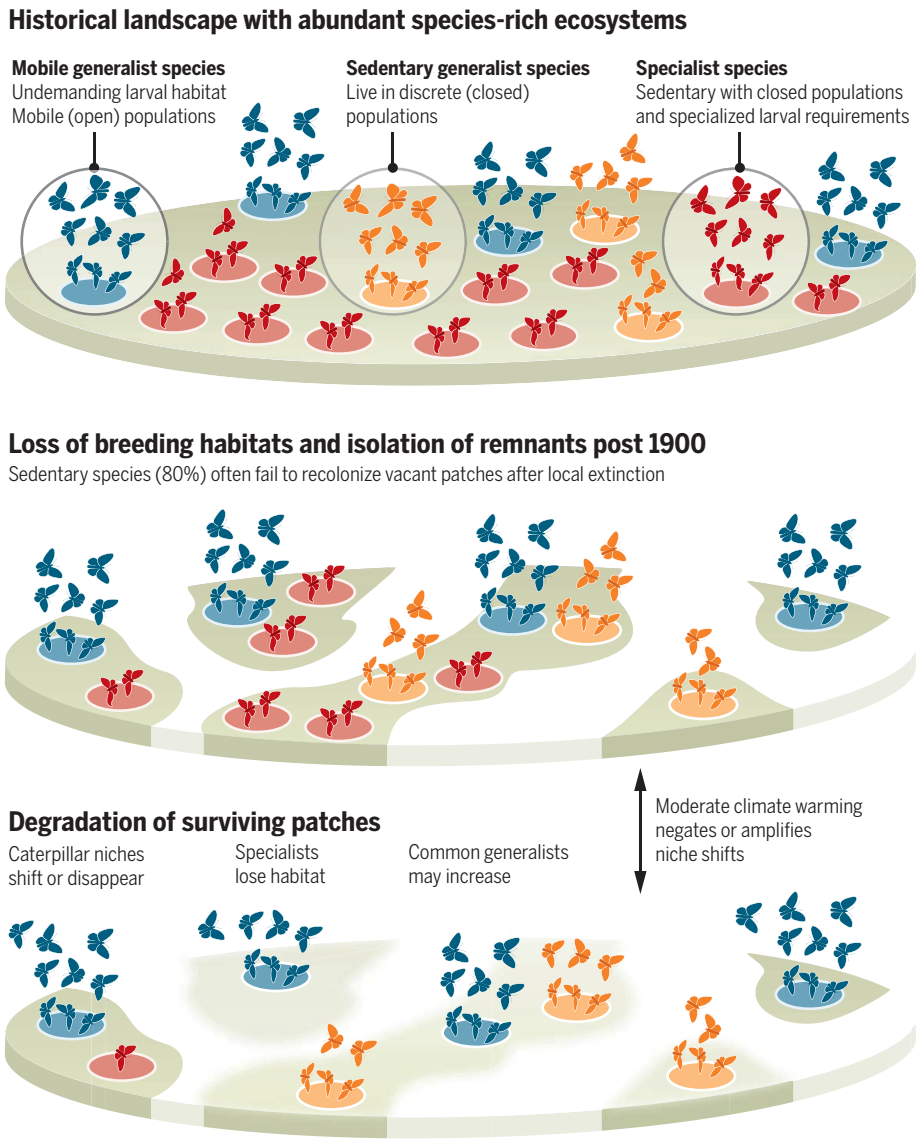
The intensification of human land use, especially for agriculture, is the main cause of butterfly declines and of changes in community composition (see the figure) (*2, 3, 6, 10, 14*). One process that drives changes in butterfly assemblages is a fundamental loss of breeding areas as essential larval food plants are replaced by crops, sown grass leys, exotic plantations, urbanization, and other land-use changes. These changes have eliminated most traditional, species-rich lowland grasslands in developed countries, a process that has started and is often advanced in developing ones. Butterflies and other taxa have, however, also declined and changed on the fragments that survive as isolated islands of seminatural habitats in modern landscapes. These shifts are attributable to two constraints in their population dynamics (*14*).

One constraint is low adult dispersal. About 80% of known butterfly species live in closed populations supported by small (often less than 2 ha), discrete patches of breeding habitat, with little migration between sites separated by more than 1 to 2 km of inhospitable ground. When populations go extinct from time to time on isolated sites—as they always have, but now do with increased frequency owing to the deteriorating quality and small size of many sites—it becomes progressively less likely that vacant or new patches will be recolonized quickly enough for a metapopulation of interlinked populations to persist in a landscape. This process leads to the disappearance of many sedentary butterflies from modern landscapes, and to a preponderance of mobile species in their depleted assemblages (see the figure).

The other constraint is the specialism of the larval habitat or niche. Whereas adult butterflies generally use wide-ranging, exchangeable resources that (the monarch apart) seldom limit their numbers, the quality of caterpillar habitats is the major determinant of population densities (*14*). Most larvae eat just one or a few related plant species, of which only a subset is palatable on most sites, defined by their nutrient quality, growth form, microclimate, or other attributes specific to the butterfly in question (*14*). Furthermore, about 25% of butterfly species interact with ants in their young stages; some require high densities of particular ant species to co-occur with the food plant. Almost every butterfly niche studied has been found to be more fastidious than had been supposed (*14*), but it is still convenient to class species into two types: specialists (~60% species) and the common wider countryside generalists. The former live mainly in closed populations, whereas roughly half of the latter are more mobile.

## HABITAT DEGRADATION

The second, less conspicuous process that drives change in butterfly assemblages is a shift or subtle degradation of larval habitats




GRAPHIC: K. SUTLIFF/*SCIENCE*

in surviving fragments of ecosystem (see the figure) (*2, 6, 14*). Many unproductive grasslands, especially hillsides, that once experienced intermittent low-intensity grazing are today abandoned as uneconomic to farm. At the same time, much woodland management has shifted to producing mature crops of (often exotic) tree species, very unlike the frequent small-scale disturbances generated by coppicing (in which trees or shrubs are periodically cut back to ground level to stimulate growth) and other defunct practices. Both processes lead to greater uniformity in the habitats available to butterfly larvae and typically to a shift from early- and mid-successional stages of vegetation to taller, denser, shadier ones. This may benefit a few common generalist species, but again at the expense of numerous specialists.

More insidious still is landscape-scale degradation from pollution (especially increased nitrogen concentrations), drainage schemes, and climate change. The last is considered to have had a largely neutral impact on butterfly populations to date, enhancing numbers and expansions in the cooler sectors of species' climatic ranges while depleting them in warmer parts (*15*). However, the increased frequency of extreme weather and projections for future climate impacts, including increased droughts, are universally harmful (*2*, *6*, *14*).

It is disappointing that many protected areas have failed to conserve their butterfly assemblages despite unchanged plant diversity (*6, 14*). A better understanding of species ecologies and of the processes that drive population changes makes it possible to restore suitable conditions. In practice, restoration often focuses on a few endangered species that nevertheless prove to be umbrellas for numerous other rarities that thrive under their treatments (*6, 16*). Thus, four of the six nationally threatened butterflies that became extinct on all or most of their UK reserves between 1960 and 1989 have now returned to a greater representation in protected areas than previously recorded (*12*), including the large blue *Maculinea arion*, an iconic habitat specialist (see the photo) (*14, 16*).

## A WORLDWIDE PROBLEM

Although the factors driving change in butterfly assemblages are best understood in Europe, there is evidence of similar processes occurring in developed and developing nations worldwide, for example, in Brazil's surviving fragments of Atlantic Forest (*7*) and in U.S. prairie grasslands (*8*). Equally disturbing is the decline of the North American monarch, *Danaus plexippus*, from 1 billion to 33 million adults in the past 25 years. This, however, is a highly atypical species, made doubly vulnerable by possessing specialized larvae, exceptional in a migrant, and by the adult trait of overwintering mostly in one vast Mexican reserve (*9*).

It is reassuring that some of the most challenging species' declines can be reversed by conservation practices informed by ecological study (*2, 6, 14, 16*), but such measures are expensive. A higher priority is to assess and protect the remaining global hotspots for butterfly diversity (*10*); to establish a comprehensive suite of more local reserves; and to understand and conserve the historical successions and dynamics of their ecosystems, be they primary forest, natural grasslands, or the low-input pastoral landscapes that survive in many developing nations, including eastern Europe (*6*). Monitoring change in butterfly assemblages is an essential first step (*4*). It is thus encouraging that the rigorous schemes in Europe are being extended to other regions such as South Africa, China, and Australia (*10*). Urgent, too, is the need for research to explore how well the drivers of declines identified in temperate regions apply to the tropics, and to assess the plasticity of butterfly phenotypes to adapt to future climatic and land-use changes. ■

***"...many protected areas have failed to conserve their butterfly assemblages despite unchanged plant diversity."***


### REFERENCES

1. O. T. Lewis, M. J. M. Senior, *J. Insect Conserv.* **15**, 121 (2011).
2. R. Fox *et al.*, *The State of the UK's Butterflies 2015* (Butterfly Conservation and Centre for Ecology and Hydrology, Wareham, 2015).
3. M. S. Warren *et al.*, *Nature* **414**, 65 (2001).
4. J. A. Thomas, *Philos. Trans. R. Soc. B Biol. Sci.* **360**, 339 (2005).
5. J. C. Habel *et al.*, *Conserv. Biol.* 10.1111/cobi.12656 (2016).
6. M. Wallis DeVries, I. Wynhoff, C. Van Swaay, Eds., http://vlindernet.nl/doc/future4butterflies_book_of_abstracts_def_2016.pdf (2016).
7. K. S. Brown Jr., A. V. L. Freitas, *Biotropica* **32**, 934 (2000).
8. S. R. Swengel *et al.*, *J. Insect Conserv.* **15**, 327 (2011).
9. S. Jepsen *et al.*, *Conservation Status and Ecology of the Monarch Butterfly in the United States* (NatureServe, Arlington, VA, and the Xerces Society for Invertebrate Conservation, Portland, OR, 2015).
10. T. R. New, Ed., *Insect Conservation: Past, Present and Prospects* (Springer, Dordrecht, Netherlands, 2012).
11. D. P. A. Sands, T. R. New, in *Conservation of the Richmond Birdwing Butterfly in Australia*, D. P. A. Sands, T. R. New, Eds. (Springer, Dordrecht, Netherlands, 2012), pp. 1–27.
12. K. F. Conrad *et al.*, *Biol. Conserv.* **132**, 279 (2006).
13. D. Baker *et al.*, *PLOS ONE* **11**, e0157423 (2016).
14. J. A. Thomas *et al.*, *J. Insect Conserv.* **15**, 241 (2011).
15. C. Parmesan *et al.*, *Nature* **399**, 579 (1999).
16. J. A. Thomas, D. J. Simcox, R. T. Clarke, *Science* **325**, 80 (2009).


10.1126/science.aaf8838

GENE EXPRESSION

# *Chromatin controls behavior*

## Dynamic regulation of chromatin remodeling controls learning and memory

*By* **J. David Sweatt**

Chromatin structure stabilizes and compacts the genome to package it within the nucleus. This structure also serves as a dynamic regulator of gene expression, silencing or activating transcription depending on molecular signals impinging upon it. It has been understood for the past two decades that chromatin stabilizes gene readout after cell-fate determination, establishing and perpetuating the precise pattern of genes transcribed in a given cell to maintain its phenotype (*1, 2*). But what about dynamic regulation of chromatin structure and its biological role? On page 300 of this issue, Yang *et al.* (*3*) describe how dynamic regulation of chromatin remodeling controls cerebellar circuit development, function, and cerebellum-dependent learning and memory, and challenge prevailing epigenetics dogma in the central nervous system.

One process that controls chromatin structure and transcription is chromatin remodeling by energy-dependent protein complexes (*4*). Such complexes that depend on adenosine 5′-triphosphate (ATP) can control gene expression by moving, ejecting, or restructuring nucleosomes, the scaffolds around which DNA wraps. Each nucleosome contains a core particle of eight histone protein subunits (*5*). Remodeling events include posttranslational modifications, swapping individual histone subunit isoforms into and out of the core particle, and facilitating the unbinding of DNA from the core particle.

The nucleosome remodeling and deacetylase (NuRD) complex catalyzes dynamic ATP-dependent nucleosome remodeling (*6, 7*). Among its constituent subunits are histone deacetylases (HDACs), chromodomain helicase DNA binding protein 3 or 4 (CHD3/4), and methyl-CpG binding domain protein 3 (MBD3). CHD4 is necessary for

*Department of Pharmacology, Vanderbilt University, Nashville, TN 37232, USA. Email: david.sweatt@vanderbilt.edu*

ATP-dependent remodeling, and was specifically targeted by Yang *et al.* in their study. In addition, the NuRD complex also enables histone subunit exchange, in which one particular histone subunit (e.g., histone H2A) can be swapped out and replaced with a different isoform (e.g., histone H2A.z) (*8*). The different histone subunit isoforms confer on their "parent" core particle different properties concerning the likelihood of the attendant DNA sequence being actively transcribed. These various processes had been extensively studied in model systems such as cultured cells, but only sparsely in more complex biological systems such as the central nervous system (*9*).

Yang *et al.* examined the intriguing idea that ATP-dependent dynamic chromatin remodeling controls neural circuit refinement and cognitive function in the cerebellum, a brain region fundamentally involved in motor control and various types of procedural learning. The authors used a multifaceted approach employing next-generation sequencing and bioinformatics, neuroanatomical visualization, in vivo cell biology and physiology, and learning behavior. As a first step, they investigated mouse cerebellar genome structure to determine the distribution of histone posttranslational modifications and the histone H2A.z subunit isoform at active cerebellar genes. Next-generation sequencing approaches and the mouse rotating rod (rotarod) behavioral motor learning task revealed that behavioral activity causes changes in gene transcription and chromatin composition, both of which depended on NuRD and ATP (see the figure).

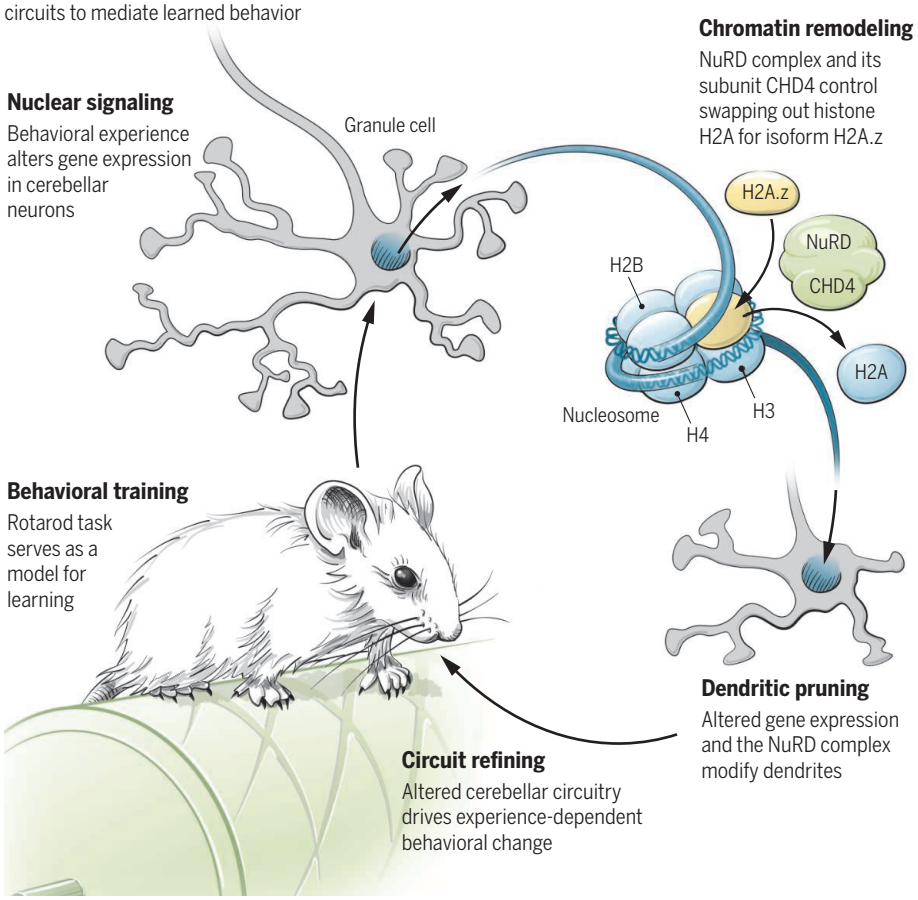
To investigate gene regulation in developmentally synchronized mouse cerebellar granule neurons, Yang *et al.* used translating ribosome affinity purification (TRAP) and discovered that the NuRD complex is necessary for activity-dependent gene regulation in these neurons in vivo. Particularly interesting was that the NuRD complex controlled granule neuron dendritic pruning. A central hypothesis regarding cerebellar circuit function is that dendritic pruning refines circuitry and enables sparse encoding of information. This concept is important to many computational models of cerebellar information processing, and Yang *et al.* confirm that NuRD-regulated dendritic pruning likely contributes to cerebellar circuit refinement in vivo.

Yang *et al.* also imaged activity (calcium signaling) in individual granule neurons in real time, in a live behaving mouse executing a rotarod task, and observed sparse encoding of granule cell firing during real-time information processing. Conditional deletion of the gene encoding CHD4 in cerebellar granule neurons in mice led to aberrant responsiveness of these neurons. These mice were hyperresponsive to normal physiologic stimulation, consistent with a disruption of sparse encoding of real-time information in animals deficient in NuRD activity. Furthermore, in the absence of CHD4, mice displayed deficits in two known cerebellum-dependent learning and memory tasks. These animals had impaired rotarod learning behavior, which ties a functional behavioral consequence (rotarod learning) back to an original stimulus for establishing behavior-dependent alterations in gene expression and chromatin occupancy. They also showed pronounced deficits in cerebellar eye-blink conditioning, a Pavlovian associative learning paradigm that is dependent on cerebellar circuit function.

The study of Yang *et al.* greatly expands our understanding of the role of dynamic regulation of the structure of the chromatin core particle in central nervous system function in the behaving animal. Moreover, the authors nicely illustrate that high-level cerebellum-dependent cognitive function hinges on normal processes underlying chromatin regulation. The findings extend recent discoveries of dynamic regulation of the previously assumed "stable" core of the chromatin particle in the central nervous system (*10-12*), and advance the fields of neuroepigenetics and neurodevelopment (*13*) by implicating an important new mechanism in transcriptional regulation of neural circuit development and refinement. Moreover, the discoveries of Yang *et al.* are an important addition to an emerging literature implying a dynamic epigenome in the central nervous system (*14, 15*), which is contrary to the prevailing dogma in the epigenetics field. ■

## Chromatin alterations mediate behavioral change

Training causes altered gene readout and plasticity of cerebellar circuits to mediate learned behavior




GRAPHIC: K. SUTLIFF/*SCIENCE*


### REFERENCES

1. A. Burton, M. E. Torres-Padilla, *Nat. Rev. Mol. Cell Biol.* **15**, 723 (2014).
2. A. Rudenko, L. H. Tsai, *Neuropharmacology* **80**, 70 (2014).
3. Y. Yang *et al.*, *Science* **353**, 300 (2016).
5. J. W. Conaway, *Annu. Rev. Biochem.* **81**, 61 (2012).
4. J. S. Runge *et al.*, *Curr. Top. Dev. Biol.* **117**, 1 (2016).
6. J. Signolet, B. Hendrich, *FEBS J.* **282**, 1692 (2015).
7. M. P. Torchy *et al.*, *Cell. Mol. Life Sci.* **72**, 2491 (2015).
8. O. Gursoy-Yuzugullu *et al.*, *J. Mol. Biol.* **428**, 1846 (2016).
9. Y. Xue *et al.*, *Mol. Cell* **2**, 851 (1998).
10. I. Maze *et al.*, *Neuron* **87**, 77 (2015).
11. I. B. Zovkic, J. D. Sweatt, *Neuron* **87**, 1 (2015).
12. I. B. Zovkic *et al.*, *Nature* **515**, 582 (2014).
13. J. D. Sweatt, *Neuron* **80**, 624 (2013).
14. E. Borrelli *et al.*, *Neuron* **60**, 961 (2008).
15. J. J. Day, J. D. Sweatt, *Nat. Neurosci.* **13**, 1319 (2010).


10.1126/science.aah4055

ECOLOGY

# How much biodiversity loss is too much?

## Widespread biodiversity losses are observed but safe-limit thresholds remain uncertain

*By* **Tom H. Oliver**

How much of something do we need to keep people safe and well? This question is frequently asked by those working in risk management. Across diverse sectors from flood protection to health care, practitioners assess risk as the product of the impact of a given event and the probability of its occurrence. Although these estimates are often uncertain, policy-makers must ultimately make spending decisions aimed at averting these risks, because the costs of inaction to society can be substantial. Biodiversity loss is a similarly critical, yet uncertain, issue. On page 288 of this issue, Newbold *et al.* (*1*) quantify global biodiversity losses, providing much-needed information on the encroachment of proposed "safe limits."

Economic analyses suggest that the total global value of ecosystem services is in the realm of tens of trillions of dollars (*2*). Many of these ecosystem services are underpinned by biodiversity. However, there is currently a lack of coordinated action to halt biodiversity declines, despite repeated setting of international targets (*3*). We know, broadly, the types of actions that are needed. They include habitat restoration as well as limiting human-derived pressures such as habitat loss, pollution, and invasive species. But the opportunity costs of these actions, in combination with the high levels of uncertainty around biodiversity change, appear to hamper commitment to action.

This uncertainty has multiple components. We must ascertain both the current extent of biodiversity losses and the effects of these losses on people's health and well-being. Newbold *et al.* report a crucial advance in tackling these issues. Their analysis is the most comprehensive quantification of global biodiversity change to date, considering over 1.8 million records of abundance from 39,123 species across 18,659 sites. Biodiversity losses vary widely across biomes. The authors find that, on average, the local abundance of each species has fallen to ~85% of its original value in the absence of human land use; that is, there is 85% "biodiversity intactness" (*4*). The authors then go further to relate these losses to a planetary safe limit of 90% biodiversity intactness, as proposed in a recent study (*5*). The hypothesis is that below the safe limit, the wide range of services provided by biodiversity that underpin human well-being—such as crop pollination, waste decomposition, regulation of the global carbon cycle, and cultural services that are central to emotional and spiritual health—are critically threatened (*5*). Newbold *et al.* find that ~58% of the world's land surface, and 9 out of 14 of the world's terrestrial biomes, have fallen below this safe threshold.

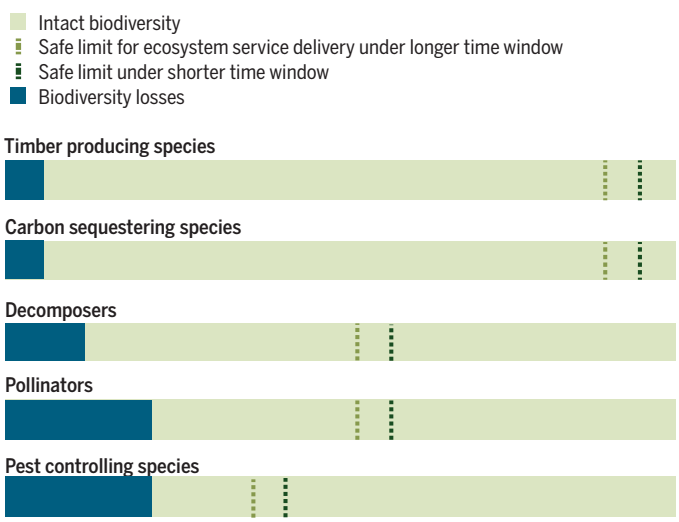
If such a large proportion of land has already passed the safe planetary boundary for biodiversity loss, why have we not already noticed more widespread negative effects on humans? Biodiversity loss can clearly lead to dramatic and rapid effects on ecosystem services. For example, invasion by the spiny water flea *Bythotrephes longimanus* in Lake Mendota in Madison, Wisconsin, USA, caused declines in key algal-grazing zooplankton species and consequent reductions in water quality, which will cost $86 million to $163 million to restore (*6*). In many other cases, however, effects may be delayed, with ecosystem services only lost after further perturbation (*7*). By analogy, cumulative structural damage to a bridge may only lead to sudden collapse after an extreme storm. Recovery from such catastrophic "tipping points" can be very costly if the replacement cost far exceeds ongoing repair costs. But the environment may be unique in that the extinction of species is essentially irreversible.

The existence of tipping points in nature has been hotly contested in the ecological literature. Debate is ongoing over whether thresholds for tipping points are planetary or regional (*8*), but they could be both or neither. It is also unclear whether non-native species should be included in the accounting of biodiversity change; it depends on their capacity to replace the roles played by disappearing native species.

To tackle these uncertainties, Newbold *et al.* conducted sensitivity analyses. When they included non-native species as functioning biodiversity, only 48% of the global land area was below the safe limit. The results differed even more markedly, however, when the threshold for safe levels of biodiversity intactness was varied. If the biodiversity intactness safe limit is 80% rather than 90%, about one-third of the global land area falls below the safe limit (and even less if non-native species are included in calculations). Clearly, it is crucial to reduce uncertainty around the existence of tipping points and where we should set safe-limit thresholds. Further research should also investigate whether safe limits are context dependent. For example, the longer the time window of

### Boundaries for biodiversity loss on services

The extent of biodiversity losses varies between groups of species that provide different services (*12*), as may the safe limits beyond which biodiversity loss will have substantial effects on human well-being. The figure only shows a subset of services provided by species. The extent of biodiversity loss and the safe limits depicted are purely hypothetical




GRAPHIC: V. ALTOUNIAN/*SCIENCE*

*University of Reading, School of Biological Sciences, Whiteknights, Harborne Building, RG6 6AS, UK. Email: t.oliver@reading.ac.uk*

interest, the higher are the chances of a large environmental perturbation, and the greater the level of system resilience needed. Safe limits may also vary between the different ecosystem services provided by species (see the figure).

Newbold *et al.*'s study marks an important step in our ability to quantify biodiversity loss. But further work is still needed. For example, the authors used statistical models to estimate local biodiversity changes by downscaling land cover changes and inferring the proportion of non-native species. Further data collection, especially in data-sparse countries, is essential to reduce uncertainty in these models. However, the most pressing knowledge gap now is to move from documenting biodiversity changes to understanding their effect on people's health and well-being. Quantitative risk assessments are urgently needed to prompt commitment to potentially costly actions. For example, the UK government only committed to action in combating climate change after the likely economic impacts were quantified through meticulous analysis combining climate science and economics (*9*).

It may be impossible, however, to reduce the uncertainty regarding consequences of biodiversity loss to levels achieved in other sectors that deal with risk management. The challenge is to take decisions in the face of this uncertainty while factoring in other social considerations, such as the fair distribution of risks across demographic groups and generations. Furthermore, some level of natural resource exploitation, with inevitable biodiversity loss, may be essential to raise the living standards of the world's poorest (*10*). Taking decisions in the face of such uncertainty is not easy. It is a tricky problem to say how much biodiversity loss is too much. However, we can be certain that inaction commits us to a future with substantial costs to human well-being (*2*, *11*). ■


## REFERENCES AND NOTES

1. T. Newbold *et al.*, *Science* **353**, 288 (2016).
2. R. Costanza *et al.*, *Global Environ. Change* **26**, 152 (2014).
3. D. P. Tittensor *et al.*, *Science* **346**, 241 (2014).
4. R. J. Scholes, R. Biggs, *Nature* **434**, 45 (2005).
5. W. Steffen *et al.*, *Science* **347**, 1259855 (2015).
6. J. R. Walsh *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 4081 (2016).
7. T. H. Oliver *et al.*, *Trends Ecol. Evol.* **30**, 673 (2015).
8. T. M. Lenton, H. T. P. Williams, *Trends Ecol. Evol.* **28**, 380 (2013).
9. N. Stern, *Stern Review on the Economics of Climate Change* (UK Treasury, 2006).
10. K. Raworth, *Oxf. Policy Prac.: Clim. Change Resilience* **8**, 1 (2012).
11. S. Díaz *et al.*, *PLOS Biol.* **4**, e277 (2006).
12. T. H. Oliver *et al.*, *Nat. Commun.* **6**, 10122 (2015).



## ACKNOWLEDGMENTS

I thank M. S. Heard, R. W. Jackson, M. D. Pagel, R. M. Sibly, and P. J. Stoett for discussion and comments.

10.1126/science.aag1712


GRAPHIC: ADAPTED FROM (6, 7) BY C. BICKEL/*SCIENCE*

FERROELECTRICS

# *Ferroelectric chalcogenides—materials at the edge*

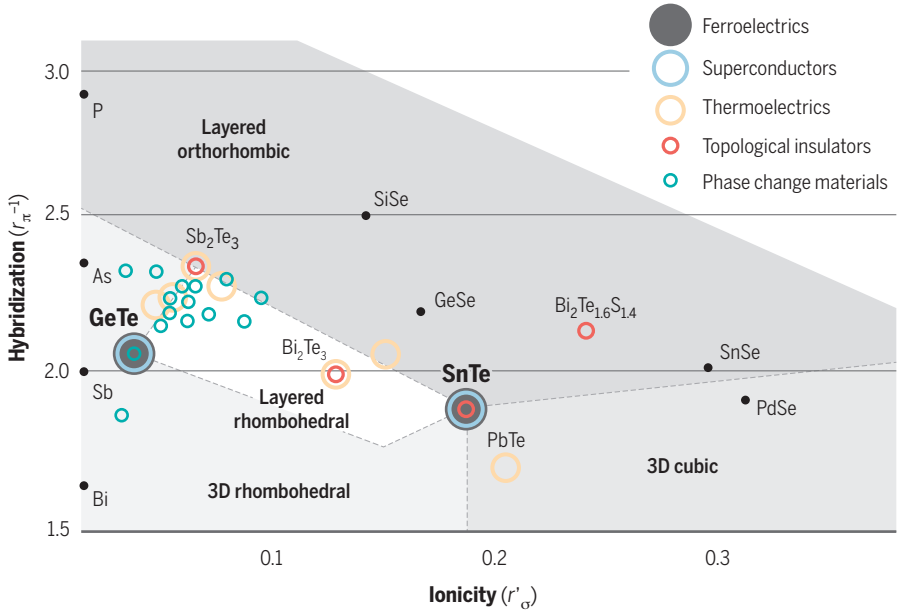
## Tin telluride becomes a more robust ferroelectric as an ultrathin film

*By* **Bart J. Kooi** *and* **Beatriz Noheda**

A ferroelectric material possesses an intrinsic electric dipole (polarization) whose direction can be reversed with an applied field. Applications of ferroelectrics include nonvolatile memories and sensors, but for high-density electronic devices or nanoscale devices, a limitation has been that as a ferroelectric film gets thinner, the maximum temperature for retaining the dipole—the Curie temperature $T_c$—decreases (often well below room temperature). On page 274 of this issue, Chang *et al.* (*1*) show that ultrathin layers of tin telluride (SnTe) can display robust, room-temperature, ferroelectric properties with higher $T_c$ than that of the bulk material.

This unusual behavior arises because SnTe and related chalcogenide materials (materials based on S, Se, or Te) can accommodate characteristics that are typically "conflicting" if the phases they form are at the edge of stability. The interplay of ionic and covalent bonding can lead to structural instabilities, so that the structure can change (for example, with temperature) between a rock-salt structure (a nonpolar atomic arrangement) and a rhombohedrally distorted structure (which is polar and can form a ferroelectric phase). Materials of this family often display asymmetric weaker and stronger bonds, in which the weak interactions can become van der Waals interactions and create layered materials that can be exploited as two-dimensional (2D) materials. These changes in bonding also underlie the variety of electronic structures in these materials, which can range from metallic to insulating.

Most chalcogenides have a clear separation between the *s*- and *p*-orbitals, and only the latter are involved in the bonding. The rock-salt structure is typically associated with ionic bonding, but GeTe, which is closely akin




**Balancing on the edge.** The type of crystal formed by chalcogenide materials and their properties depend on the interplay between *s*-*p*-orbital hybridization (reflecting covalent bonding with the amount of overlap between *s*- and *p*-orbitals) and ionicity (reflecting their tendency to form a salt) (*6*, *7*). Only a couple of these materials are known to be ferroelectric at low temperatures: GeTe, which is at the stability edge between two crystal phases, and SnTe, which is at the edge corner between all the known chalcogenide phases. Chang *et al.* show that few-atoms-thick films of this material stabilize the ferroelectric phase even at room temperature, which make such SnTe films interesting for real devices. [The figure is based on the hybridization versus ionicity plots of (*6*, *7*).]

to SnTe, has a rock-salt structure dominated by covalent bonding (see the figure). Because only three *p* electrons are present to stabilize the six bonds, the concept of resonance bonding was introduced to explain the observed structures (*2*). This resonance bonding makes the material's properties susceptible to relatively small perturbations, so potentially, many properties can be switched.

For the well-known perovskite ferroelectrics, such as $PbTiO_3$, $BaTiO_3$, or $BiFeO_3$, the temperature $T_c$ at which intrinsic polarization is lost and above which the material becomes a paraelectric decreases when the film thickness is reduced. However, Chang *et al.* observed that in SnTe, $T_c$ is substantially higher for thin films than for the bulk material, even surpassing room temperature for a film with a thickness of only two unit cells. Achieving stable ferroelectric macroscopic polarization at room temperature in ultrathin films is usually a great challenge because the depolarization fields increase as film thicknesses decrease. Such in-plane polarized films that are robust against depolarization fields hold great promise for memory applications. From the fundamental point of view, the results are also very important because they show a direct image of the band-bending induced by the presence of ferroelectric polarization, as well as the magnitude of the associated screening length.

***"Indeed, a rich physics has been theoretically predicted to arise from the coexistence of ferroelectricity and spin-orbit coupling..."***

According to Chang *et al.*, the enhancement of $T_c$ arises from quantum confinement effects. Although SnTe is an ionic 3D crystal, the few layers of SnTe deposited on a graphitized silicon carbide surface are only weakly bound to the substrate via van der Waals interactions and form a truly 2D crystal. Thus, these layers are very different from regular epitaxially grown ferroelectric layers of perovskite materials. The quantum confinement across the film (along the *c* axis of the unit cell) leads to a huge increase in the band gap from 0.8 eV (in bulk) to above 1.5 eV (in a single unit-cell layer). The greater band gap reduces the density of conduction-band charge carriers, which reduces the screening of the Coulomb dipolar interactions and stabilizes the ferroelectric phase.

*Zernike Institute for Advanced Materials, University of Groningen, Netherlands. E-mail: b.j.kooi@rug.nl; b.noheda@rug.nl*

This reduction of carriers with respect to the bulk material is also achieved through the increase of the Sn-vacancy formation energy in the thin films. Indeed, Sn vacancies in bulk SnTe cause it to be a heavily *p*-doped semiconductor. However, the density of vacancies in these as-grown ultrathin SnTe layers is four orders of magnitude lower than in bulk. Chang *et al.* argue that the in-plane lattice expansion (and the associated polar distortions) taking place through a well-known lattice relaxation mechanism in the surface layers of rock-salt ionic crystals also help stabilize in-plane ferroelectric distortions.

Although these factors are playing important roles, it is most likely not the whole story behind the origin of ferroelectricity in SnTe. In the related material GeTe, the rock-salt–to-rhombohedral phase transition is associated with a Peierls distortion (a lowering in energy by decreasing symmetry). The half-filled *p*-band, in combination with the symmetry breaking along one of the four <111> directions, opens a band gap and reduces the energy (*3*). Thus, it might be worth investigating whether a similar transition in SnTe drives symmetry breaking. Interestingly, the ultrathin layers of SnTe form domains, which is unexpected in such thin, in-plane polarized films that are not strongly bound to the substrate. In addition, in-plane polarization along the <110> directions is only compatible with the orthorhombic symmetry, which would bring 2D SnTe closer to the more ionic SnSe (see the figure).

The distinctive properties of SnTe and similar chalcogenides are governed by a fine balance between atomic interactions that can develop very different characters (ionic, covalent, and resonant bonding). These different bonding motifs not only allow ample tuning of the electronic band structure and defect energy formation but also bring these materials to the boundary between different structural instabilities (including ferroelectric phases) and to the edge between 2D and 3D behavior. Indeed, a rich physics has been theoretically predicted to arise from the coexistence of ferroelectricity and spin-orbit coupling, a key characteristic of topological insulators based on chalcogenides such as $Bi_2Se_3$ and $Sb_2Te_3$ (*4*, *5*). ■

## REFERENCES


1. K. Chang *et al.*, *Science* **353**, 274 (2016).
2. G. Lucovsky, R. M. White, *Phys. Rev. B* **8**, 660 (1973).
3. P. B. Littlewood, *Crit. Rev. Solid State Mater. Sci.* **11**, 229 (1983).
4. E. Plekhanov, P. Barone, D. Di Sante, S. Picozzi, *Phys. Rev. B* **90**, 161108(R) (2014).
5. U. V. Waghmare *et al.*, *Phys. Rev. B* **67**, 125111 (2003).
6. M. Wuttig *et al.*, *Allg. Chem.* **638**, 2455 (2012).
7. P. B. Littlewood, *J. Phys. C Solid State Phys.* **13**, 4855 (1980).


10.1126/science.aaf9081

COGNITION

# *Thinking abstractly like a duck(ling)*

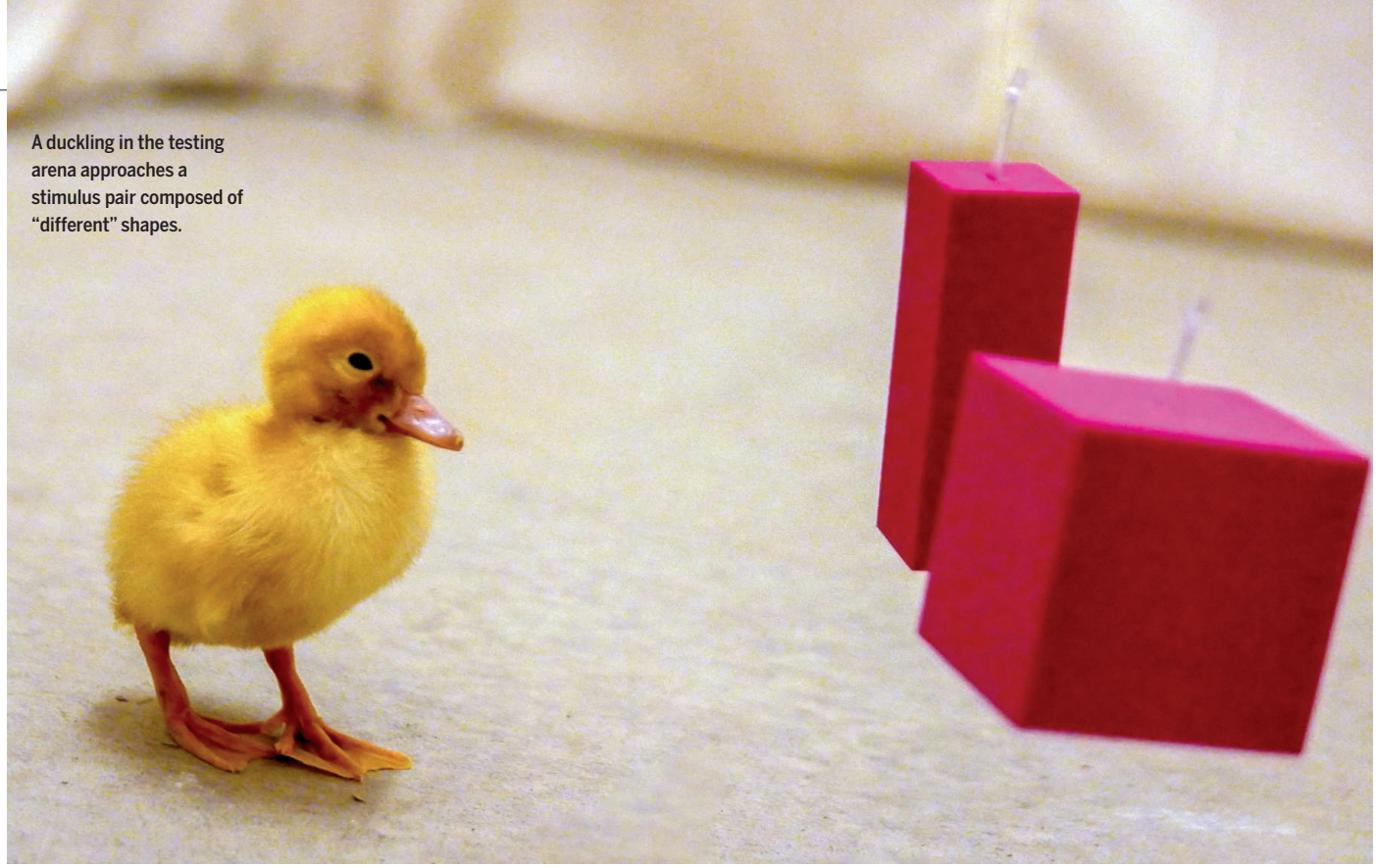
## Ducklings, like human infants, show signs of abstract conceptual thought

*By* **Edward A. Wasserman**

In the film *O Brother, Where Art Thou?*, when one of a trio of bungling prison escapees angrily asks another, "Who elected you leader of this outfit?" his buddy smugly quips, "I figured it should be the one with the capacity for abstract thought." Indeed, abstract conceptual thought is held to be so central to being human that the idea of someone being incapable of this kind of thinking is a subject for (sometimes rather cruel) humor. Interest in understanding the capacity for abstract thought has been a matter of serious consideration that dates back at least three centuries to the famous English philosopher John Locke. Locke confidently contended that "brutes abstract not" (*1*) and insisted that exhibiting abstract thought definitively divided humans from all other animals. However, no science then existed to confirm or refute Locke's contention. On page 286 of this issue, Martinho and Kacelnik (*2*) put the claim that animals are incapable of abstract thought to a strong behavioral test.

Adhering to the adage that "actions speak more loudly than words," scientists are deploying powerful behavioral tests that provide animals with nonverbal ways to reveal their intelligence to us. Although animals may not be able to speak, studying their behavior may be a suitable substitute for assaying their thoughts, and this in turn may allow us to jettison the stale canard that thought without language is impossible. Following this behavioral approach and using the familiar social learning phenomenon of imprinting, Martinho and Kacelnik report that mallard ducklings preferentially followed a novel pair of objects that conformed to the learned relation between a familiar pair of objects. Ducklings that had earlier been exposed to a pair of identical objects pre-

*Department of Psychological and Brain Sciences, DeLTA Center, University of Iowa, Iowa City, IA 52242, USA. Email: ed-wasserman@uiowa.edu*



A duckling in the testing arena approaches a stimulus pair composed of "different" shapes.

PHOTO: ANTONE MARTINHO

ferred a pair of identical objects to a pair of nonidentical objects; other ducklings that had been exposed to a pair of nonidentical objects preferred a pair of nonidentical objects to a pair of identical objects. Because the testing objects were decidedly unlike the training objects, Martinho and Kacelnik concluded that the ducklings had effectively understood the abstract concepts of "same" and "different."

This study is important for at least three reasons. First, it indicates that animals not generally believed to be especially intelligent are capable of abstract thought. Second, even very young animals may be able to display behavioral signs of abstract thinking. And, third, reliable behavioral signs of abstract relational thinking can be obtained without deploying explicit reward-and-punishment procedures.

Discriminating same from different sets of stimuli is a fundamental cognitive skill that now appears to be within the ability of many different species of animals. Not surprisingly, primates—including chimpanzees (*3*), rhesus monkeys (*4*), and baboons (*5*)—have exhibited this ability when given a wide range of behavioral tasks. So too have less obviously intelligent creatures such as rats (*6*) and several species of birds, including pigeons (*7*), parrots (*8*), and crows (*9*). In addition, same-different concept learning has been reported in honey bees (*10*) and bumblebees (*11*). The claim that abstract relational thinking is a unique ability of human beings can no longer be supported.

It is also noteworthy that newly hatched ducklings show clear signs of abstract conceptual thought. The same is true of human infants. Even 7-month-olds can acquire the general concepts of same and different, although it is necessary for them to see those relations exemplified by four distinct pairs of stimuli (*12*). The ducklings in the Martinho and Kacelnik study needed to see those relations exemplified by only a single pair of stimuli.

The study with human infants (*12*) used a habituation design in which looking time was used to assess abstract relational thinking. After exposure to a pair of either identical or nonidentical stimuli, infants preferentially looked at unfamiliar pairs of stimuli that involved the opposite stimulus relation—a novelty preference. This habituation technique and the imprinting methodology used in the Martinho and Kacelnik study are both cases of unsupervised learning; no "carrot-and-stick" methods were necessary to encourage the organisms to exhibit experimenter-demanded performance—so-called supervised learning. The fact that abstract relational concepts can be acquired and expressed without deploying explicit reward-and-punishment procedures further extends the ecological and adaptive importance of abstract relational learning.

Human adults surely possess an extremely powerful capacity for abstract relational cognition. Yet mounting behavioral evidence suggests that there is both ontogenetic and phylogenetic continuity in the nature of abstract thought. Human infants and nonhuman animals may exhibit similar basic relational abilities. However, adult humans come to dramatically outperform both infants and other animals in the complexity and intricacy of their abstract thinking. Humans' experience with language and culture probably promotes this advantage. Nevertheless, we should appreciate that whatever heights of cognition may be attainable by humans must have arisen via an evolutionary process about which we may gain key insights by studying the cognitive processes of our animal kin.

There is every reason to expect future research to delve more deeply into the species generality of abstract thought and to more fully explore the cognitive and neural mechanisms of relational thinking. Similar behaviors may or may not reflect common processes. ■

## REFERENCES


1. J. Locke, *An Essay Concerning Human Understanding* (Oxford, 1690).
2. A. Martinho III, A. Kacelnik, *Science* **353**, 286 (2016).
3. D. Premack, *Intelligence in Ape and Man* (Erlbaum, 1976).
4. B. M. Basile, E. J. Moylan, D. P. Charles, E. A. Murray, *Anim. Cogn.* **18**, 1221 (2015).
5. E. A. Wasserman, J. Fagot, M. E. Young, *J. Comp. Psychol.* **115**, 42 (2001).
6. E. A. Wasserman *et al.*, *Learn. Memory* **19**, 142 (2012).
7. E. A. Wasserman, M. E. Young, *J. Exp. Psychol. Anim. B* **36**, 3 (2010).
8. T. Obozova, A. Smirnova, Z. Zorina, E. Wasserman, *Anim. Cogn.* **18**, 1363 (2015).
9. A. Smirnova *et al.*, *Curr. Biol.* **25**, 256 (2015).
10. M. F. Brown, J. M. Sayde, *Anim. Cogn.* **16**, 117 (2013).
11. M. Giurfa *et al.*, *Nature* **410**, 930 (2001).
12. A. L. Ferry *et al.*, *Child Dev.* **86**, 1386 (2015).


10.1126/science.aag3088

POLICY FORUM

SOCIAL MEDIA RESEARCH

# *Crisis informatics—New data for extraordinary times*

## Focus on behaviors, not on fetishizing social media tools

*By* **Leysia Palen**[1,2*] *and* **Kenneth M. Anderson**[1]

Crisis informatics is a multidisciplinary field combining computing and social science knowledge of disasters; its central tenet is that people use personal information and communication technology to respond to disaster in creative ways to cope with uncertainty. We study and develop computational support for collection and sociobehavioral analysis of online participation (i.e., tweets and Facebook posts) to address challenges in disaster warning, response, and recovery. Because such data are rarely tidy, we offer lessons—learned the hard way, as we have made every mistake described below—with respect to the opportunities and limitations of social media research on crisis events.

### SOCIAL MEDIA CAN BE FETISHIZED

Too much importance is attributed to social media as a tool instead of to the behaviors that underlie it. At the same time, too little attention is granted to examining the "corners" of social media spaces where interesting forms of work and coalitions of helpers are found, e.g., to help people struggling in the recovery after the 2010 Haiti earthquake by "topping off" their phones with minutes (*1*) and offering language translation of texts (*2*), or after Hurricane Sandy with pet-family reunion (*3*). These creative efforts are important but include only a relatively small number of people in contrast to the loud and voluminous general social media response that arises after major disasters.

Yet in the world of emergency management, we often see a demand for science that "proves" behaviors. For example, does the use of social media save lives or improve emergency management? One wonders if the same questions were asked about the first landline telephones. Traditionally, surveys have been used to ask people after the fact how they received warning and evacuation information. Such research on social media use is problematic. The surveys are issued with some delay after the event. People cannot readily recall where different pieces of information came from. Furthermore, the distinction between where and from whom we learned the information is misplaced; e.g., the sheriff is a person, an agency, and a digital presence.

Underlying this survey research is both doubt and fetishization of social media: Are people really using it? Should we invest money in communicating this way? To appreciate that social media are being used in disasters can in turn mean radical changes must occur to emergency management organizations. Such overhaul comes at a cost. Hence there is caution and request for "proof."

Although doubt about social media's value by emergency managers is understandable, crisis informatics researchers question whether such proof is attainable and if it can be used meaningfully in emergency management decisions in an evolving information environment. We argue that such proof is impossible and that demand by members of the public during an event is what will change the course of emergency management even if "best practices" dictate otherwise (*4, 5*).

Investigation should shift from the burden of proof of use to design in use. Enduring social media solutions for emergency management have yet to be discovered. What is necessary is to have sufficient permission by emergency management to support solutions as they emerge from grassroots operations and then to foster those ideas deliberately in subsequent events (*2, 5*).

For example, the Humanitarian OpenStreetMap Team (HOT) movement has drawn upon volunteer mappers to make geospatial data open and available in response to the 2010 Haiti earthquake, 2013 Typhoon Yolanda, and the 2015 Nepal earthquakes. International responders have grown to appreciate the value of the data. Although HOT strives to make data and process improvements in anticipation of big events, responders know that sometimes gains in collaborative mapping practice happen during the event (*6*).

Data scientists are relatively new to the disaster–social media research space and are knowledgeable about the management and analytics of large-volume data. But their knowledge and, hence, their questions commonly focus on the highest-level correlations that come about when comparing available large data sets [e.g. (*7*)]. This work is welcome and relevant but is only the beginning of a set of questions that need to be asked about behavioral phenomena.

In principle, social science should figure strongly in social media disaster research. Social science research recognizes often ignored, but critical, distinctions between human behaviors that result from endogenous hazards that can be captured (e.g., crime-based events) versus exogenous hazards that cannot (e.g., weather events that give rise to natural disasters) (*5*). In social media research, these differences are often flattened, even though they lead to different interpretations and outcomes. Marginalization of these critical fields in data science is worrisome: inclusion of social science would allow for more robust computational social science (*8*).

### SAMPLING DECISIONS ARE DIFFICULT

A tempting myth is that large volumes of social media data alone will reveal patterns of

A funnel cloud near Venice, Italy, June 2012.

PHOTO: PETER PHIPP/TRAVELSHOTS.COM/ALAMY STOCK PHOTO

[1]*Department of Computer Science, University of Colorado, Boulder, CO 80309, USA.* [2]*Department of Information Science, University of Colorado, Boulder, CO 80309, USA.*
**Email: leysia.palen@colorado.edu*

behavior, when in truth, data must be scoped in ways that make it pertinent to a meaningful question, which usually triggers a cascade of new data collection steps and questions (*9, 10*). When the focus is on volume, rigor in data collection becomes an afterthought. We argue that social media research is the same as any scientific endeavor with respect to the relationship between data collection and accurate research questions; the abundance of data disguises this obligation.

For example, our collaborators at the National Center for Atmospheric Research helped identify keywords to locate tweets about Hurricane Sandy before landfall on 29 October 2012. Keywords had to be broad ("frankenstorm" and "sandy") because we did not know where the storm would land nor what the important research questions would be. As the place of landfall became clear, we added localized terms, e.g., place names. Collecting on broad terms like #sandy samples the "global" population of curious onlookers, but for sampling the "local" population, which shares and seeks information differently, one has to rely on highly localized terms. Geotagged tweets can help determine a user's proximity to an event; however, only 1 to 2% of tweets are geotagged.

> ***"Social science research recognizes often ignored, but critical, distinctions between human behaviors [in response to] hazards...."***

Even this is not enough to enable many questions, because, in return, one receives only a tweet: It represents just one time a place was mentioned or a tweet was geotagged. What are users saying in between such tagged tweets? Are there meaningful data there? When we collect on specialized terms about a disaster, we get data skewed toward the disaster. How can we counterbalance such bias? Our solution includes collecting "contextual streams" on people identified first in the coarse-grained, keyword–based search. In our initial keyword–collected data set for Sandy, we identified 92,000 users who used geotagging on the U.S. eastern seaboard. We then retrieved their most recent 3200 tweets (a limitation imposed by Twitter) to generate a data set of 205 million tweets. Then, we could compare their historical data (with respect to geotagged movement) with their movement immediately before, during, and after the hurricane. This "boring in" on an interesting problem—in this case, evacuation or shelter-in-place movement—is where the power of social media lies with respect to social science problems.

## MAKE "BIG" DATA BIGGER, THEN SMALLER

Social media are inherently about participation. Social media data do not necessarily represent all of a population evenly, but they do represent a range of behaviors, ideas, and opinions that have a role to play alongside traditional disaster response. Disaster response can be better served if we plan to accept and allow people to have a voice and to help in disaster response.

In addition to this lofty call, social media data provide the opportunity for new insights on a larger scale. For example, instead of being in place to witness actions by members of the public, social media reveals some of what is happening on the ground at sites worldwide. It can be observed with a different kind of precision, including temporal sequencing of events and discourse that can be analyzed without transcription. The trick, though, is to be able to approach the "big data" qualitatively and even ethnographically.

To isolate activity by location or with respect to new and unusual behaviors, data sets must get bigger (by collecting contextual streams), before they can be sampled or filtered accordingly. This is because there are few natural constraints on social media data. There is no automatic mechanism for drawing one's "unit of analysis" and scope. The bounds of observation must be done through decisions—which may have acknowledged limitations—to scope the data. Once made smaller, content analysis helps determine whether the decisions were reasonable; if not, then back to the drawing board. In this way, researchers can isolate populations, regions, and small groups talking about or taking action, e.g., the Far Rockaway community in New York, where many residents felt underserved during Hurricane Sandy (*11*).

## THE TYRANNY OF THE TWEET

Researchers and emergency managers interested in social media are at the mercy of social media providers (*12*) to gain access to data. Even when data are available, we are limited by the data delivery format. For example, Twitter makes data available in the JavaScript Object Notation (JSON) format, with each JSON object containing a single tweet without its conversational context. Yet we know that people speak with continuity across their posts without repeating terms that are most likely to be keywords (*13–18*). As a result, most research studies tweets in isolation as single statements without the monologic context, never mind the conversational context. This, we find, renders most of them useless in terms of understanding their value. But, in context, they are often enlightening. Furthermore, it is difficult to reacquire that context as it requires substantial postprocessing, e.g., to expand a data set 10-fold to search for those contextual tweets. In this way, the data format influences how researchers conceptualize what can be done.

We have offered these lessons not just to caution but also to compel cross-disciplinary research in a critical and exciting area of societal import. The pursuit of deeper investigations enabled by comprehensive treatment of social media data can reveal surprising and emergent features of disaster response in a technologically mediated world.

## REFERENCES AND NOTES


1. K. Starbird, L. Palen, in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, Vancouver, BC, 7 to 12 May 2011 [Association for Computing Machinery (ACM), New York, 2011], pp. 1071–1080.
2. P. Meier, *Digital Humanitarians: How Big Data Is Changing the Face of Humanitarian Response* (CRC Press, Boca Raton, FL, 2015).
3. J. White, L. Palen, K. M. Anderson, in *Proceedings of ACM Conference on Computer Supported Cooperative Work*, Baltimore, February 2014 (ACM, New York, 2014), pp. 866–876.
4. A. L, Hughes, L. A. A. St. Denis, L. Palen, K. M. Anderson, in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, Toronto, 26 April to 1 May 2014 (ACM, New York, 2014), pp. 1505–1514.
5. L. Palen, "Frontiers in crisis informatics," invited talk at the ACM Conference on Human Factors in Computing Systems, Seoul, South Korea, 20 April 2015 (ACM, New York, 2015).
6. R. Soden, L. Palen, in *International Conference on the Design of Cooperative Systems*, Nice, France, 27 to 30 May 2014 (Springer, New York, 2014), pp. 311–326.
7. Y. Kryvasheyeu *et al.*, *Science Adv.* **2**, 3 (2016).
8. A. Zelenkauskaite, E. P. Bucy, *First Monday* **21**, 5 (2016).
9. C. Castillo, *Big Crisis Data: Social Media in Disasters and Time-Critical Situations* (Cambridge Univ. Press, Cambridge, 2016).
10. K. M. Anderson, A. Schram, A. Alzabarah, L. Palen, *IEEE Bull. Tech. Com. Data Eng.* **36**, 3 (2013).
11. J. Anderson *et al.*, in *Proceedings of the Information Systems for Crisis Response and Management*, Rio de Janeiro, Brazil, 22 to 25 May 2016 (Information Systems for Crisis Response and Management, 2016), 16 pp.
12. D. Lazer *et al.*, *Science* **343**, 1203 (2014).
13. L. Palen *et al.*, in *ACM and BCS Visions of Computer Science* Edinburgh, UK, 13 to 16 April 2010 (ACM, New York, and British Computing Society, Swinton, UK, 2010), art. 8.
14. M. Kogan, L. Palen, K. M. Anderson, in *Proceedings of the ACM Conference on Computer-Supported Cooperative Work*, Vancouver, BC, 14 to 18 March 2015 (ACM, New York, 2015), pp. 981–993.
15. K. Starbird, L. Palen, A. Hughes, S. Vieweg, in *Proceedings of the ACM Conference on Computer-Supported Cooperative Work*, Savannah, GA, 6 to 10 February 2010 (ACM, New York, 2010), pp. 241–250.
16. K. M. Anderson, A. Schram, in *Proceedings of the International Conference on Software Engineering*, Waikiki, Honolulu, HI, 21 to 28 May 2011 (ACM, New York, 2011), pp. 844–847.
17. A. Sarcevic *et al.*, in *Proceedings of the ACM Conference on Computer-Supported Cooperative Work*, Bellevue, WA, 11 to 15 February 2012 (ACM, 2012), pp. 47–56.
18. K. Glasgow in *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, Ann Arbor, MI, 1 to 4 June 2014 (Association for the Advancement of Artificial Intelligence, Palo Alto, CA, 2014), pp. 161–169.


## ACKNOWLEDGMENTS

This material is based on work sponsored by NSF grants AGS-1331490, IIS-1524806, and IIS-0910586. We thank all who have been affiliated with Project EPIC: Empowering the Public with Information in Crisis.

10.1126/science.aag2579



Jim and Eileen Matthews walk along the fenced-in Love Canal site in Niagara Falls, NY in 2004.

PHOTO: DAVID DUPREY/ASSOICATED PRESS

ENVIRONMENTAL POLICY

# *A toxic timeline*

## Lessons from Love Canal offer context for complacent environmental policies

*By* **Jacob Darwin Hamblin**

In the 1970s, residents of a Niagara Falls neighborhood realized that chemicals from a toxic waste dump had leached into their homes, parks, and neighborhood school. Their cancers, miscarriages, and myriad chronic ailments told the tale, and in 1978 they organized, filed lawsuits, and demanded intervention. The federal government eventually complied, evacuating the residents and creating the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA), also known as the Superfund Act, which provides a framework for cleaning up such sites.

Similar events have played out in towns across America, from Hinkley, California, to Flint, Michigan. It seems like a performance piece finely tuned to the conditions of our era. Yet, in his new book, Richard S. Newman urges us to see the Love Canal disaster stretched out in time, rooted in the long history of the Niagara Falls area. The crisis itself, he says, was an outcome of patterns established generations earlier that pitted developmental pressures against environmental and human health and created a "cycle of disposable land use that had long dominated area politics and economics." The Love Canal residents of the 1970s, he says, were just the first to successfully challenge it.

After surveying the Niagara region in the late 19th century, William Love imagined a hydroelectric canal to power a great manufacturing city but left it only partly dug when funding for the project evaporated. "Love saw the landscape itself as a great canvas upon which to draw monumental scenes," Newman writes. "When the picture faded, Love simply moved on, leaving to subsequent generations the task of reckoning with environmental impacts he set in motion."

Soon after Love departed, Elon Hooker, an engineer turned entrepreneur, arrived. Hooker Electrochemical Company's midcentury "Chemagination" campaign positioned the industry as a great emancipator of humanity, facing the perils of industrialization by cleaning the air and water with chemistry. Between 1939 and 1950, its profits rose 850% as it supplied disinfectants, DDT, materials for wartime explosives, and many other goods to the military and postwar consumers. Hooker began putting toxic sludge into William Love's abandoned waterway in 1942.

Faced with a burgeoning population, local officials set their sights on the parcel of land that contained Love's canal in the 1950s. When the Niagara Falls school board bought the site from Hooker for one dollar in 1953, the company unmistakably warned about the dangers associated with it. Yet the board appeared hell-bent on building a new school. The foundation of the 99th Street School sank into the chemical pit during construction. It was moved 30 yards north and opened in 1955. When the board sold some of the land to a developer to build homes, Hooker once again warned against digging foundations. The houses were built anyway.

Throughout Newman's narrative is a palpable sense that citizens must remain active and vigilant in the face of government complacency. In the 1970s, the Love Canal Homeowners Association, along with the Concerned Love Canal Renters Association, transformed residents into activists clamoring for health protection, evacuation, and compensation. But even after President Jimmy Carter declared an emergency in 1978 and Congress passed the Superfund Act in 1980, residents distrusted the authorities and constantly challenged statements made by state and federal officials.

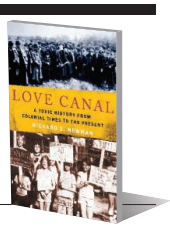
Ironically, the successful evacuation of residents, achieved under the auspices of the Superfund Act, removed the very source of activism that made the protests successful. Taking a leaf out of Hooker's playbook, the city paid organizations to recast Love Canal as a place cleansed by the marvels of science and technology. In 1990, 10 refurbished houses were reopened for sale and the neighborhood was repackaged as Black Creek Village. Attracted by the area's affordability and location, prospective homeowners expressed confidence in the Superfund site's remediation. Many activists and former residents remain skeptical.

*Love Canal* challenges readers to think about long-term structural problems that are place-specific and deeply historical. That is a laudable aim, although the search for commonalities in the "very long history of environmental protest along the Niagara Frontier" leads Newman to use "environmental" capaciously, including, for example, to describe the colonial disputes between Native Americans and white settlers. Nevertheless, he succeeds in revealing the public health fiasco as a powerful example of persistent citizen activism in the face of government complacency.

The term "Love Canal" has become a metaphor for a particular kind of public health crisis that could happen in Anytown, USA. Newman asks us to remember too that Love Canal was also a specific place in a specific time and that the human costs of development have long been recognized and resisted.

*The author is at the School of History, Philosophy, and Religion, Oregon State University, Corvallis, OR 97330, USA. Email: jacob.hamblin@oregonstate.edu*

**Love Canal**
**A Toxic History from Colonial Times to the Present**
*Richard S. Newman*
Oxford University Press, 2016.
306 pp.




10.1126/science.aag0454

HEALTH ANALYTICS

# Big data meets human health

## Internet searches and fitness trackers are poised to play a role in the future of health care

*By* **Conor Farrington**

Last fall, I visited Stanford University to present new research on an artificial pancreas system for people with type 1 diabetes. Although I don't have diabetes myself, I chose to wear the system. I had never monitored my glucose levels before, and it was fascinating—and at times alarming—to see how they fluctuated in response to food, drink, stress, fatigue, exercise, and sleep. I had unconsciously joined a large and growing community of self-trackers.

As Gina Neff and Dawn Nafus emphasize in their excellent book *Self-Tracking*, there is a "veritable explosion" in the number of people using digital and wearable devices to record, analyze, and reflect upon data created by their own bodies and behaviors. Self-tracking activities can be undertaken for a wide range of motivations—for example, to maintain wellness, monitor specific medical conditions, foment political change, or create new forms of digital art and fashion. Data artist Laurie Frick, for example, has developed FrickBits, an app that converts smartphone GPS data into abstract collages, and sculptor Stephen Cartwright creates elegant representations of his tracked movements through space and time.

These activities depend on the latest technology, but Neff and Nafus point out that self-tracking has a long history. Benjamin Franklin used charts to monitor his productivity and moral character, giving particular attention each week to one of 13 virtues.

But can self-tracking induce positive change for the average contemporary user? Neff and Nafus quote an exchange from Stanford Medicine X 2012, an annual conference that explores the potential roles that emerging technologies will play in the future of medicine and health care. Here, one panelist argued, "Data leads to knowledge and knowledge leads to change." If this were true, "there would be no need for psychologists," countered another panelist. And indeed, although there is some evidence that self-tracking data can induce behavioral changes, social scientists have convincingly demonstrated the motivational and data interpretation challenges that confront self-trackers.

Neff and Nafus tread carefully between the twin pitfalls of techno-utopianism and techno-dystopianism to develop a nuanced position that acknowledges both the opportunities and the challenges raised by self-tracking. In doing so, they raise a number of important questions. Can quantitative self-tracking data fully capture who we are as individuals? Will self-tracking practices continue to exclude the sickest and poorest in our societies? In medical contexts, will self-tracking patients manage the burden of interpreting their own data, and will health care professionals be willing to incorporate such data into more personalized care? Who profits financially from self-tracking data? And, more broadly, will we continue to defend the right to data privacy, or will our societies shift toward total transparency?

A visitor demos a mobile blood-pressure monitor at the Wearable Expo in Tokyo.

PHOTO: BLOOMBERG/CONTRIBUTOR

Elad Yom-Tov's *Crowdsourced Health* discusses a different field of digital health, in which data are generated not through the use of wearable devices but from queries entered in search engines. Based on the premise that these searches mirror our offline behavior and that the Internet offers greater privacy and accessibility than many other possible sources of information, he suggests that these data could reveal information about health that would be difficult or impossible to gather in other ways.

Yom-Tov, a researcher at Microsoft Research, describes the studies he has carried out on medical conditions, ranging from anorexia to bipolar disorder. Among the more striking assertions is his claim to have generated a quantitative foundation for the Kübler-Ross "five stages of grief" model. Using cancer diagnosis as a proxy for bereavement, Yom-Tov and his team tracked queries and page visits by cancer patients and their families. They separated the most-visited sites into 11 categories and used a hidden Markov model to reveal the five predicted stages. They also revealed different progressions through the stages, depending on whether the searcher's cancer was acute or chronic, leading Yom-Tov to suggest that a more tailored communication method might be of benefit to patients.

Yom-Tov also claims to have discovered a new class of long-term drug side effects by tracing search queries about mild symptoms. While serious side effects emerging soon after drug use begins are likely to be reported to the FDA, he argues that patients may not associate long-term and milder symptoms with specific drugs, thus hindering their discovery as side effects.

*Crowdsourced Health* draws an overly stark divide between traditional medical research and newer digital research approaches, in part because Yom-Tov overlooks the contributions of qualitative and mixed-methods research in medicine. And for a book focused on Big Data, relatively little detail about the author's methods of analysis appears in the text. Nevertheless, Yom-Tov succeeds in demonstrating that online queries represent a potentially important source of medical data.

The question that has yet to be answered is what should ultimately be done with all of these data—and by whom.


**Self-Tracking**
*Gina Neff and Dawn Nafus*
MIT Press, 2016. 233 pp.

**Crowdsourced Health How What You Do on the Internet Will Improve Medicine**
*Elad Yom-Tov*
MIT Press, 2016. 144 pp.

*The reviewer is at the Cambridge Centre for Health Services Research, University of Cambridge, Cambridge CB2 0SR, UK. Email: cjtf2@medschl.cam.ac.uk*

10.1126/science.aaf8769

# LETTERS

*Edited by* **Jennifer Sills**

## Mining undermining Brazil's environment

ON 5 NOVEMBER 2015, two mining dams jointly owned by Vale and BHP Billiton collapsed in Brazil. The toxic sludge wiped out whole villages, leaving 19 dead and suffocating 600 km of one of Brazil's most valuable rivers (*1*). One might expect increased scrutiny on the mining sector's social responsibilities after such an event. Instead, Brazilian politicians—many of whom owe their campaign funding to mining companies—have recently pushed forward three pieces of legislation intended to neuter environmental regulations.

The first bill (PL 37/2011) will create the National Mining Agency, giving more autonomy to the sector. In effect, it will allow the mining industry to decide which areas to mine. A complementary piece of legislation (PL 1610/1996) will authorize mining in indigenous lands. There are 4181 official requests from mining companies that are held up by current regulations and awaiting the vote on this bill. If it passes, this will open the possibility for these companies to operate in 177 indigenous territories (*2*). Indigenous lands cover one-fifth of the Brazilian Amazon and currently represent one of the most effective barriers against deforestation (*3*). Finally, the constitutional amendment PEC 65/2012 aims to weaken the licensing process of large developments by eliminating the current power that environmental agencies have to suspend a project based on its Environmental Impact Assessment (*4*, *5*). In practice, Environmental Impact Assessments will be reduced to a tick-box requirement for the mining industry.

The biomes most threatened by mining—Amazon, Atlantic Forest, and Cerrado (*6*)—are also the most valuable in terms of biodiversity and ecosystem services (*7*). Approval of these three pieces of legislation would signify that Brazil is no longer at the forefront of sustainability and confirm that its politics are vulnerable to development pressures and corporate lobby (*6*). We agree with Edwards and Laurance (*8*) that an independent global authority to police large multinational mining companies in developing nations is urgently needed.

***Hani Rocha El Bizri,[1]* Jonathan Christopher Bausch Macedo,[2] Adriano Pereira Paglia,[3] Thaís Queiroz Morcatty[4]***

[1]Programa de Pós-Graduação em Saúde e Produção Animal na Amazônia, Universidade Federal Rural da Amazônia, Belém, Pará, Brazil. [2]Instituto de Desenvolvimento Sustentável Mamirauá, Tefé, Amazonas, Brazil. [3]Laboratório de Ecologia e Conservação, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil. [4]Wildlife Conservation Society–Brazil, Manaus, Amazonas, Brazil.

*Corresponding author. Email: hanibiz@gmail.com

**REFERENCES AND NOTES**

1. H. Escobar, *Science* **350**, 1138 (2015).
2. A. Almeida, S. Futada, T. Klein, "Protected areas and indigenous lands in the Amazon region are affected by more than 17,500 mining processes," Instituto Socioambiental (2016); www.socioambiental.org/en/node/5044 [in Portuguese].
3. D. Nepstad *et al.*, *Conserv. Biol.* **20**, 65 (2006).
4. L. Wade, *Science* **352**, 1044 (2016).
5. *Science* received this Letter on 13 May 2016, before the News story by L. Wade (*3*) was written. Wade used the Letter as background for the story.
6. J. Ferreira *et al.*, *Science* **346**, 706 (2014).
7. N. Myers *et al.*, *Nature* **403**, 853 (2000).
8. D. P. Edwards, W. F. Laurance, *Science* **350**, 1482 (2015).

10.1126/science.aag1111

## Brazil's Amazon conservation in peril

IN HER NEWS In Depth story "Brazilian crisis threatens science and environment" (27 May, p. 1044), L. Wade explored the consequences of the controversial constitutional amendment that is under consideration in the Brazilian Senate amidst the country's economic and political turmoil. This amendment, known as PEC 65/2012, effectively abolishes Brazil's environmental licensing process (*1*). Ratification could lead to large-scale, indiscriminate destruction of the Amazon biome.

About 334 dams have been proposed throughout the Amazon basin, and more than half of them are in the Brazilian Amazon (*2*). In addition, over 1 million square kilometers of the Brazilian Amazon have been registered as under consideration for mining (*3*). The implementation of many of these proposed projects is hindered by environmental restrictions: Sixty percent of the Amazonian hydropower potential (*4*) and 20% of Amazonian areas with registered interest for mining (*3*) lie inside strictly protected areas and indigenous lands. If ratified, the new amendment will allow developers to ignore environmental restrictions.

Brazil's most recent 10-year energy expansion plan states that 12 megadams must be completed in the Amazon basin by 2024. These dams represent 93% of the country's projected increase in hydropower generation capacity (*5*). If the amendment is ratified, these future dams—together with associated infrastructure megaprojects such as highways and electricity transmission lines—would be implemented despite insufficient impact assessment. If the associated highway construction also lacks sustainable planning, the dam project could indirectly lead to an indiscriminate expansion of agricultural frontiers and an increase in deforestation rates.

To protect the Amazon ecosystem, we must modernize Amazon energy plans, replace conventional infrastructure with sustainable infrastructure, and integrate planning and management. Recent calls for basin-scale planning before new infrastructure projects (*2*, *3*) are constructed are on the right track. The proposed amendment would derail these plans, putting decades of conservation efforts and the Amazon system itself in jeopardy.

***Rafael M. Almeida,[1,2]* Thomas E. Lovejoy,[3] Fábio Roland[1]***

[1]Department of Biology, Federal University of Juiz de Fora, Juiz de Fora, 36015-440, Brazil. [2]Cary Institute of Ecosystem Studies, Millbrook, NY 12545, USA. [3]Department of Environmental Science and Policy, George Mason University, Fairfax, VA 22030, USA.

*Corresponding author. Email: rafaelmarquesjf@yahoo.com.br

**REFERENCES**

1. Proposta de Emenda à Constituição N° 65 de 2012 (http://legis.senado.leg.br/mateweb/arquivos/mate-pdf/120446.pdf) [in Portuguese].
2. K. O. Winemiller *et al.*, *Science* **351**, 128 (2016).

Proposed changes to regulations would allow mining projects in Brazil's indigenous territories.


3. J. Ferreira *et al.*, *Science* **346**, 706 (2014).
4. Brasil Ministério de Minas e Energia (MME), Plano Nacional de Energia 2030 (MME, Rio de Janeiro, 2007) [in Portuguese].
5. Brasil Ministério de Minas e Energia (MME), Plano Decenal de Expansão de Energia 2024 (MME, Rio de Janeiro, 2014) [in Portuguese].


10.1126/science.aag2510

# *Brazil's Amazonian fish at risk by decree*

A MISGUIDED POLICY, based not on science but rather on the need to generate profit, threatens the highly diverse native fish communities in the Amazon River Basin. A controversial law—4330/2016 (*1*)—signed by the governor of Amazonas State, Brazil, allows the rearing of non-native fish and genetically modified species, as well as the damming of streams for this purpose, even in Areas of Permanent Protection.

In the midst of a Brazilian political crisis (*2*), the law was sanctioned without consulting the federal government, environmental agencies, or the public. The ratification of the law was not approved by the Ministry of Environment, which opposes it (*3*). The Amazon Basin hosts the most species-rich fish fauna in the world (*4*), and this fauna remains uncontaminated by non-native fishes (*5*). The law is dangerous because one of the main drivers of species introduction in Brazil is the escape of fish from fish farming facilities (*6*). Escape of non-native species with high invasive capacity is a serious threat to aquatic biodiversity (*7*, *8*), a threat that can be exacerbated if the escaped fish are genetically modified.

Moreover, aquaculture with non-native species in the Amazon Basin is unnecessary given that native fisheries can be profitable and provide socioeconomic benefits without putting the native fauna at risk. Finally, the new law contradicts current federal laws 5197/67 (*9*) and 9605/98 (*10*), which prohibit and criminalize the introduction of species.

***R. M. Tófoli,[1,2]* G. H. Z. Alves,[1,3] R. M. Dias,[1,3] L. C. Gomes[1]***

[1]Programa de Pós-Graduação em Ecologia de Ambientes Aquáticos Continentais (PEA), Universidade Estadual de Maringá, Maringá, PR, Brazil. [2]Faculdade Ingá (Uningá), Maringá, PR, Brazil. [3]Postdoctoral Fellowship, CNPq, Universidade Estadual de Maringá, Maringá, PR, Brazil.

*Corresponding author. Email: biotofoli@gmail.com


## REFERENCES


1. Assembléia Legislativa do Estado do Amazonas (2016); http://legislador.aleam.gov.br/LegisladorWEB/LegisladorWEB.ASP?WCI=ProjetoTexto&ID=201&INEspecie=1&nrProjeto=79&aaProjeto=2016 [in Portuguese].
2. L. Wade, *Science* **352**, 1044 (2016).
3. "MMA quer revogação de lei que permite peixes não-nativos no estado do Amazonas" (2016); www.mma.gov.br/index.php/comunicacao/agencia-informma?view=blog&id=1635 [in Portuguese].
4. W. J. Junk *et al.*, *Aquat. Ecosyst. Health. Manag.* **10**, 2(2007).
5. V. M. Azevedo-Santos *et al.*, *Nat. Conserv.* **13,** 2 (2015).
6. J. C. G. Ortega *et al.*, *Hydrobiologia* **746**, 1 (2015).
7. F. M. Pelicice *et al.*, *Conserv. Lett.* **7**, 1 (2014).
8. S. C. Fornek *et al.*, *Hydrobiologia* **1**, 773 (2016).
9. Código de Caça - Lei ° 5197/67 | Lei N° 5.197, de 3 de janeiro de 1967 (1967); http://presrepublica.jusbrasil.com.br/legislacao/103429/codigo-de-caca-lei-5197-67 [in Portuguese].
10. Lei de Crimes Ambientais - Lei 9605/98 | Lei n° 9.605, de 12 de fevereiro de 1998 (1998); http://presrepublica.jusbrasil.com.br/legislacao/104091/lei-de-crimes-ambientais-lei-9605-98 [in Portuguese].


10.1126/science.aag2922

## TECHNICAL COMMENT ABSTRACTS

### Comment on "Sensitivity of seafloor bathymetry to climate-driven fluctuations in mid-ocean ridge magma supply"

***Maya Tolstoy***


Olive *et al.* (Reports, 16 October 2015, p. 310) and Goff (Technical Comment, 4 September 2015, p. 1065) raise important concerns with respect to recent findings of Milankovitch cycles in seafloor bathymetry. However, their results inherently support that the Southern East Pacific Rise is the optimum place to look for such signals and, in fact, models match those observations quite closely.


Full text at http://dx.doi.org/10.1126/science.aaf0625

### Response to Comment on "Sensitivity of seafloor bathymetry to climate-driven fluctuations in mid-ocean ridge magma supply"

***J.-A. Olive, M. D. Behn, G. Ito, W. R. Buck, J. Escartín, S. Howell***


Tolstoy reports the existence of a characteristic 100,000-year period in the bathymetry of fast-spreading seafloor but does not argue that sea level change is a first-order control on seafloor morphology worldwide. Upon evaluating the overlap between tectonic and Milankovitch periodicities across spreading rates, we reemphasize that fast-spreading ridges are the best potential recorders of a sea level signature in seafloor bathymetry.


Full text at http://dx.doi.org/10.1126/science.aaf2022

## ERRATA

**Erratum for the Report "Genomic correlates of response to CTLA-4 blockade in metastatic melanoma" by E. M. Van Allen *et al.*,** *Science* **352**, aaf8264 (2016). Published online 15 April 2016; 10.1126/science.aaf8264

**Erratum for the Research Article "Architecture of the type IVa pilus machine" by Y.-W. Chang *et al.*,** *Science* **352**, aaf7977 (2016). Published online 8 April 2016; 10.1126/science.aaf7977

**Erratum for the Report "MYC regulates the antitumor immune response through CD47 and PD-L1" by S. C. Casey *et al.*,** *Science* **352**, aaf7984 (2016). Published online 8 April 2016; 10.1126/science.aaf7984

TECHNICAL COMMENT

OCEANOGRAPHY

# Comment on "Sensitivity of seafloor bathymetry to climate-driven fluctuations in mid-ocean ridge magma supply"

Maya Tolstoy

**Olive *et al*. (Reports, 16 October 2015, p. 310) and Goff (Technical Comment, 4 September 2015, p. 1065) raise important concerns with respect to recent findings of Milankovitch cycles in seafloor bathymetry. However, their results inherently support that the Southern East Pacific Rise is the optimum place to look for such signals and, in fact, models match those observations quite closely.**

Olive *et al*. (*1*) and Goff (*2*) raise questions about whether recent papers (*3*, *4*) really demonstrate that "the bathymetric fabric of the seafloor…records rapid…fluctuations in ridge magma supply caused by sea level changes" (*1*) because of the importance of unrelated faulting processes. However, their criticisms are applicable primarily to Crowley *et al*. (*4*), who specifically hypothesize that "abyssal hills...record the magmatic response to changes in sea level" with a model describing static topographic compensation to changes in melt input. Tolstoy (*3*), on the other hand, hypothesizes that "Seafloor eruption rates and mantle melting fueling eruptions may be influenced by sea level and crustal loading cycles at scales from fortnightly to 100 kyr," with the primary hypothesis of the paper being that this may feed into long-term climate cycles, regardless of the bathymetric consequences. To test this, Tolstoy looked for supporting evidence in bathymetry from the ultrafast-spreading Southern East Pacific Rise (SEPR) specifically because it "provides a site where faulting is least dominant and magmatism is most prevalent" (*3*), in consideration of the complex factors contributing to seafloor bathymetry. Models of Olive *et al*. in fact demonstrate that any sea level–related signal is more likely to be observable at faster spreading rates, and Goff (*2*) leaves open the possibility that such a signal might be observed at "axial high ridges" because a subtle increase in observed widths of abyssal hill spacing with spreading rate cannot be ruled out for faster spreading rates. It is important to note that observed fault spacing represents a lower limit—i.e., the observed ~1- to 3-km spacing at fast spreading rates [figure 3A in (*1*)] (*2*, *5*) could easily mask a ~7- to 8-km spacing [~100 thousand years (ky)] superimposed on the closer spacing.

Olive *et al*. present three possible mechanisms for topographic expression of fluctuations in melt supply at the mid-ocean ridge: (i) static topographic compensation [favored by (*4*)], (ii) volcanic extrusion on the seafloor, and (iii) tectonomagmatic interactions during normal fault growth [the latter two favored by (*3*)]. Both topographic compensation and normal fault growth are shown not to be viable at Milankovitch frequencies for the intermediate spreading rates of the Australian-Antarctic Ridge (*4*), and any increased volcanic extrusion would not be at a scale that could explain the observations. However, what is not emphasized in Olive *et al*. is that all three mechanisms considered show that the most likely place to observe them is at the SEPR because of the thinner axial lithosphere, smaller faults, and faster spreading rate, with all three mechanisms capable of producing observable signals at ultrafast spreading rates (8 cm/year) depending on various assumptions: Static topography signals are modeled to potentially be as high as 30 to 50 m, although this assumed the thinnest possible end member for lithospheric strength; excess volcanism may yield an additional 40 to 55 m of topography at ~100-ky periodicity, although again this may be decreased given off-axis flow; and fault spacings for a fluctuating dike injection rate are not estimated at SEPR spreading rates, but the model shows a ~100-ky cutoff in sensitivity for intermediate spreading rates, which means that at SEPR spreading rates this cutoff would fall below 100 ky. Therefore, ~100-ky fault spacing could develop at SEPR spreading rates in the context of very strong ~100-ky modulations in melt supply.

Overall, the scales of the signals predicted at the SEPR (10s of meters and faulting) are strikingly similar to those observed (*3*) (Fig. 1). The SEPR therefore meets the conditions set by Olive *et al*.'s models for observing 100-ky periodicity through climatically driven fluctuations. They do not consider the possibility of a direct influence of eccentricity on melt production and eruption frequency (*3*), but any addition from this process would only further strengthen the case. They also do not consider the possibility of caldera collapse effects (*6*) after periods of high magmatism, which could result in the troughlike features observed during periods of correspondingly low $CO_2$ emissions (Fig. 1.).

Olive *et al*. provide an excellent starting point for discussing the likely presence and magnitude of any Milankovitch forced seafloor periodicities; however, seafloor volcanism and faulting are full of complexities that have the potential to mask such signals. For instance, at the faster spreading rates, eruptions can lead to lava flowing at least 2 km off-axis; off-axis volcanism may contribute to surface topography; and the robustness of magmatism can be highly variable along axis, the foci of which may or may not be steady through time [e.g., (*6*, *7*)]. Present-day fast-spreading ridges display areas of inflated ridge axis that may erupt more frequently and other segments that have not shown signs of an eruption for decades (*8*). As such, the approach used in Tolstoy (*3*) was to stack bathymetric

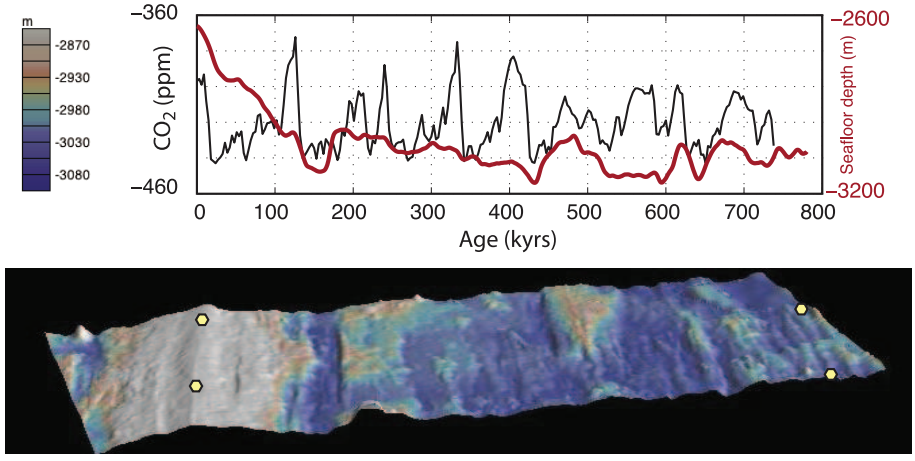



**Fig. 1. Topographic data at the SEPR compared with $CO_2$ (*3*).** (**Top**) Unfiltered stacked bathymetric profile show in red, with $CO_2$ (9) record in black. (**Bottom**) Map of area of nine axis-perpendicular bathymetric profiles (within the box shown by yellow dots) used to make the stacked profile (*10*).

Department of Earth and Environmental Sciences, Lamont-Doherty Earth Observatory of Columbia University, Palisades, NY, USA.
Email: tolstoy@ldeo.columbia.edu

profiles, a classic seismic technique for enhancing a noisy signal. In the context of all the possible causes of variability, the temporal match of an unfiltered stack of seafloor bathymetry with periods of high and low $CO_2$ output (*9*) (which roughly tracks sea level) is notable (Fig. 1).

It is worth considering that the lower-than-predicted present-day eruption rates (*3*), consistent with a period of present-day decreased magmatism, have implications for modeled mid-ocean ridge properties that are generally assumed to be steady state. For instance, data on accretion widths and axial lithospheric strength incorporated into models may not reflect the properties during a time of more robust magmatism. Systematic increases in global mid-ocean ridge volcanism may thus lead slower spreading rates to be more susceptible to the imprint of volcanic fluctuations. More data will help confirm or refute this.

The hypothesis that seafloor bathymetry is controlled by Milankovitch cycles (*4*) should not be conflated (*1*, *2*) with the hypothesis that sea level and orbital variations influence magmatism at ridges (*3*). Independent of bathymetric signatures, the consequences of non–steady state mid-ocean ridge eruptions should be considered in geochemical exchanges between the ocean and the lithosphere, including the global carbon cycle (*3*).

## REFERENCES AND NOTES

1. J.-A. Olive *et al.*, *Science* **350**, 310–313 (2015).
2. J. A. Goff, *Science* **349**, 1065 (2015).
3. M. Tolstoy, *Geophys. Res. Lett.* **42**, 1346–1351 (2015).
4. J. W. Crowley, R. F. Katz, P. Huybers, C. H. Langmuir, S.-H. Park, *Science* **347**, 1237–1240 (2015).
5. J. A. Goff, Y. Ma, A. Shah, J. R. Cochran, J.-C. Sempere, *J. Geophys. Res.* **102**, 15521–15534 (1997).
6. M.-H. Cormier *et al.*, *Geology* **31**, 633–636 (2003).
7. K. C. Macdonald, D. S. Scheirer, S. M. Carbotte, *Science* **253**, 986–994 (1991).
8. K. C. Macdonald, P. J. Fox, *Earth Planet. Sci. Lett.* **88**, 119–131 (1988).
9. D. Lüthi *et al.*, *Nature* **453**, 379–382 (2008).
10. GeoMapApp, Marine Geoscience Data System; www.geomapapp.org.

## ACKNOWLEDGMENTS

The author thanks J.-A. Olive for numerous helpful and thoughtful conversations. This work was supported by NSF under grant OCE-0961594

11 December 2015; accepted 8 June 2016
10.1126/science.aaf0625

TECHNICAL RESPONSE

OCEANOGRAPHY

# Response to Comment on "Sensitivity of seafloor bathymetry to climate-driven fluctuations in mid-ocean ridge magma supply"

J.-A. Olive,[1]* M. D. Behn,[2] G. Ito,[3] W. R. Buck,[1] J. Escartín,[4] S. Howell[3]

**Tolstoy reports the existence of a characteristic 100 thousand year (ky) period in the bathymetry of fast-spreading seafloor but does not argue that sea level change is a first-order control on seafloor morphology worldwide. Upon evaluating the overlap between tectonic and Milankovitch periodicities across spreading rates, we reemphasize that fast-spreading ridges are the best potential recorders of a sea level signature in seafloor bathymetry.**

We acknowledge the clear distinction made by Tolstoy (*1*) between the ideas put forward in (*2*) and the hypothesis central to (*3*). In (*3*), it is proposed that seafloor relief is caused primarily by changes in the height of volcanic constructions on Milankovich time scales, which reflect a control of sea level cycles on mid-ocean ridge (MOR) magma supply. Meanwhile, (*1*, *2*) point out that seafloor eruption rates may be extremely sensitive to small external perturbations such as orbital stresses. In particular, they suggest that—in addition to sea level cycles modulating overall MOR magma supply—eccentricity cycles may modulate rates of volcanic extrusion on the seafloor. A 100-thousand-year (ky) characteristic periodicity in the bathymetry of the fast-spreading Southern East Pacific Rise (SEPR) is presented to support this idea. In (*2*), however, Tolstoy does not imply that such processes are the primary cause for the abyssal hill fabric of global seafloor, which was the focus of our initial study (*4*).

A characteristic periodicity of 100 ky at the SEPR, with a full spreading rate of 14.7 cm/year, translates into a seafloor length scale of 7.4 km, which is considerably larger than the characteristic ~1- to 3-km spacing of normal faults measured at fast-spreading ridges (*5*, *6*). The biggest tectonic scarps (~50 m) in such settings correspond to axis facing faults that accommodate the small fraction of plate separation not taken up by crustal emplacement (*7*, *8*). Additional, more closely spaced faults form in response to flexural stresses as the lithosphere migrates away from the axial high, but these generally produce smaller scarps (≤20 m) (*5*, *9*–*11*). Overall, the spectral signature of the tectonic fabric of fast-spreading seafloor should consist of several distributed peaks at periodicities near and below ~40 ky (Fig. 1). Such peaks are present in the bathymetric spectrum of the SEPR but are much less energetic than the 100-ky peak (*2*). This 100-ky peak could therefore be the signature of a periodic (but nontectonic) process capable of creating coherent relief on top of the fault-induced fabric.

Based on these observations of fault geometry and distribution, our models suggest that topography-building processes at MORs act as low-pass filters—they only record fluctuations in melt supply on time scales longer than the cutoff period set by intrinsic properties of the lithosphere and magmatic accretion zone. Tolstoy rightfully points out that this cutoff period generally decreases with spreading rate, making fast-spreading ridges the best candidates to efficiently record a climatic modulation in seafloor bathymetry. Our models for volcanic extrusion specifically predict that a 10% modulation in magma supply over a 100-ky period could lead to ~50 m of relief at the SEPR, provided that the off-axis extent of lava flows remains limited. This relief would be on the order of the tectonic relief, although further modeling is needed to determine whether this modulation is sufficient to overshadow the tectonic signal in the Fourier domain.

Tolstoy proposes additional sources of relief beyond a modulation of volcanic constructions. In particular, regular episodes of summit trough collapse after periods of frequent eruptions (and associated drainage of the axial melt lens) could periodically imprint vertical offsets in seafloor bathymetry, potentially as high as ~50 m (*12*). We consider this process more plausible than a magmatic modulation of the growth of stretching faults, which would require fluctuations in magma supply an order of magnitude greater than those expected from the effect of sea level cycles (*4*, *8*). In fact, it is noteworthy that while faster-spreading ridges should constitute the most efficient topographic recorder of climate-related fluctuations in magma supply, the associated faster mantle upwelling rates are expected to damp the relative amplitude of this fluctuation (*3*). This is because the modulation in crustal production roughly scales as the upwelling rate plus the sea level–induced mantle decompression rate, normalized by the spreading rate (*13*).

Finally, we reemphasize that the complexity of the bathymetric signal should motivate additional observations that are more readily interpretable in terms of MOR magma supply. For example, oscillations in the topography of the crust-mantle boundary on 7-km length scales should be resolvable by modern multichannel seismic imaging and could be unambiguously linked to 100-ky fluctuations in crustal production at fast-spreading ridges.

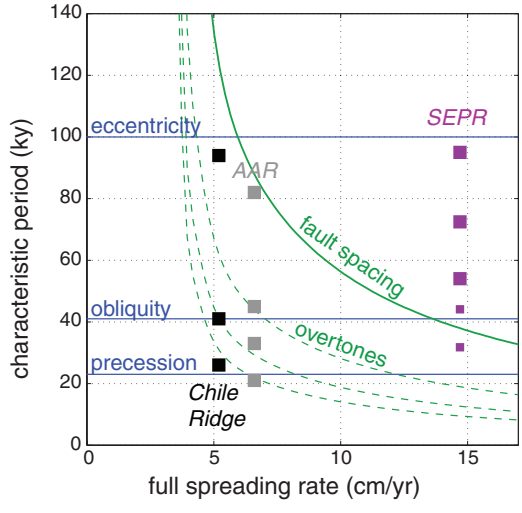



**Fig. 1. A global seafloor spectrogram.** Period of the strongest peaks in published spectra of seafloor bathymetry plotted against full spreading rate. Data come from cross-axis bathymetry spectra at the Australian-Antarctic Ridge (AAR), gray squares (*3*); Chile Ridge (CR), black squares (*14*); and the SEPR, purple squares (*2*), with two additional small squares showing small-amplitude peaks at ~30 and 44 ky). The solid green curve marks the characteristic period corresponding to the dominant fault spacing, as predicted by models that closely fit the global trend of abyssal hill spacing [green curve in figure 3A in (*4*)]. Dashed green curves indicate overtones of the dominant fault periodicity that may arise in the frequency domain (*15*). Milankovitch periodicities are reported as horizontal blue lines.

[1]Lamont-Doherty Earth Observatory, Columbia University, Palisades NY, USA. [2]Woods Hole Oceanographic Institution, Woods Hole MA, USA. [3]University of Hawaii, Honolulu HI, USA. [4]CNRS, Institut de Physique du Globe de Paris, Paris, France.
*Corresponding author. Email: jaolive@ldeo.columbia.edu

## REFERENCES AND NOTES

1. M. Tolstoy, *Science* **353**, 229 (2016).
2. M. Tolstoy, *Geophys. Res. Lett.* **42**, 1346–1351 (2015).
3. J. W. Crowley, R. F. Katz, P. Huybers, C. H. Langmuir, S. H. Park, *Science* **347**, 1237–1240 (2015).
4. J.-A. Olive *et al.*, *Science* **350**, 310–313 (2015).
5. S. M. Carbotte, K. C. Macdonald, *J. Geophys. Res.* **99** (B7), 13609–13631 (1994).
6. J. A. Goff, Y. Ma, A. Shah, J. R. Cochran, J.-C. Sempéré, *J. Geophys. Res.* **102** (B7), 15521–15534 (1997).
7. M. D. Behn, G. Ito, *Geochem. Geophys. Geosyst.* **9**, Q08010 (2008).
8. G. Ito, M. D. Behn, *Geochem. Geophys. Geosyst.* **9**, Q08010 (2008).
9. D. R. Bohnenstiehl, S. M. Carbotte, *Geochem. Geophys. Geosyst.* **2**, GC000156 (2001).
10. J. Escartín *et al.*, *Geochem. Geophys. Geosyst.* **8**, Q06005 (2007).
11. W. R. Buck, L. L. Lavier, A. N. B. Poliakov, *Nature* **434**, 719–723 (2005).
12. M.-H. Cormier *et al.*, *Geology* **31**, 633–636 (2003).
13. D. C. Lund, P. D. Asimow, *Geochem. Geophys. Geosyst.* **12**, 12009 (2011).
14. P. Huybers *et al.*, *Science* **352**, 1405 (2016).
15. J.-A. Olive *et al.*, *Science* **352**, 1405 (2016).



## ACKNOWLEDGMENTS

Funding was provided by NSF grants OCE-1154238 (M.D.B), OCE-1155098 (G.I. and S.H.), EAR-1009839 (W.R.B), CNRS support to J.E., and an LDEO Postdoctoral Fellowship for J.-A.O.

29 January 2016; accepted 8 June 2016
10.1126/science.aaf2022

# NATURE'S



CREDIT: (C) A. T. WILLETT/ALAMY STOCK PHOTO

# FURY

*By* **Brent Grocholski** *and* **Robert Coontz**

Tornadoes sweep through central Kansas. Mudslides bury a neighborhood in Guatemala. A tsunami triggers a nuclear meltdown at a power station in Japan. Every day, the news brings fresh reminders of the great dangers our planet can unleash with little warning. The dangers remain, even as researchers slowly unpack the secrets behind volcanic eruptions, earthquakes, tropical storms, and other natural hazards. That knowledge can reveal new threats, such as the planet-changing possibility of volcanic supereruptions or collisions with mountain-sized bodies from space. But it also leads to better tools for studying natural hazards and estimating where, how often, and how fiercely they are likely to strike.

Our increasingly sophisticated toolbox is helping us understand the physical processes driving hazards such as large and damaging earthquakes; track tropical cyclones to project how increasing temperatures might change these powerful storms; provide much-needed warnings of volcanoes and tsunamis; and recover from catastrophic events when they occur.

This collection of Reviews, News features, and special online material at http://scim.ag/29qQGyL explores some of the ways people are learning to assess risks, lessen dangers, and repair the damage from disasters that elude or breach our defenses. Natural hazards will never go away, but we can always become better prepared for the inevitable.



A coiling tornado looms over the central plains of Kansas—a stark symbol of the many natural dangers that remain beyond human control.



# THINKING THE UNTHINKABLE

## Rare cataclysms are hard to study and plan for, but they may be too dangerous to ignore

*By* **Julia Rosen**



Published by AAAS

Gosses Bluff meteor crater in Australia formed when a 1-kilometer-wide space rock struck Earth 142 million years ago.

PHOTO: © STEPHEN ALVAREZ/NATIONAL GEOGRAPHIC CREATIVE

In a dingy apartment building, insulated by layers of hanging rugs, the last family on Earth huddles around a fire, melting a pot of oxygen. Ripped from the sun's warmth by a rogue dark star, the planet has been exiled to the cold outer reaches of the solar system. The lone clan of survivors must venture out into the endless night to harvest frozen atmospheric gases that have piled up like snow.

As end-of-humanity scenarios go, that bleak vision from Fritz Leiber's 1951 short story "A Pail of Air" is a fairly remote possibility. Scholars who ponder such things think a self-induced catastrophe such as nuclear war or a bioengineered pandemic is most likely to do us in. However, a number of other extreme natural hazards—including threats from space and geologic upheavals here on Earth—could still derail life as we know it, unraveling advanced civilization, wiping out billions of people, or potentially even exterminating our species.

Yet there's been surprisingly little research on the subject, says Anders Sandberg, a catastrophe researcher at the University of Oxford's Future of Humanity Institute in the United Kingdom. Last he checked, "there are more papers about dung beetle reproduction than human extinction," he says. "We might have our priorities slightly wrong."

Frequent, moderately severe disasters such as earthquakes attract far more funding than low-probability apocalyptic ones. Prejudice may also be at work; for instance, scientists who pioneered studies of asteroid and comet impacts complained about confronting a pervasive "giggle factor." Consciously or unconsciously, Sandberg says, many researchers consider catastrophic risks the province of fiction or fantasy—not serious science.

A handful of researchers, however, persist in thinking the unthinkable. With enough knowledge and proper planning, they say, it's possible to prepare for—or in some cases prevent—rare but devastating natural disasters. Giggle all you want, but the survival of human civilization could be at stake.

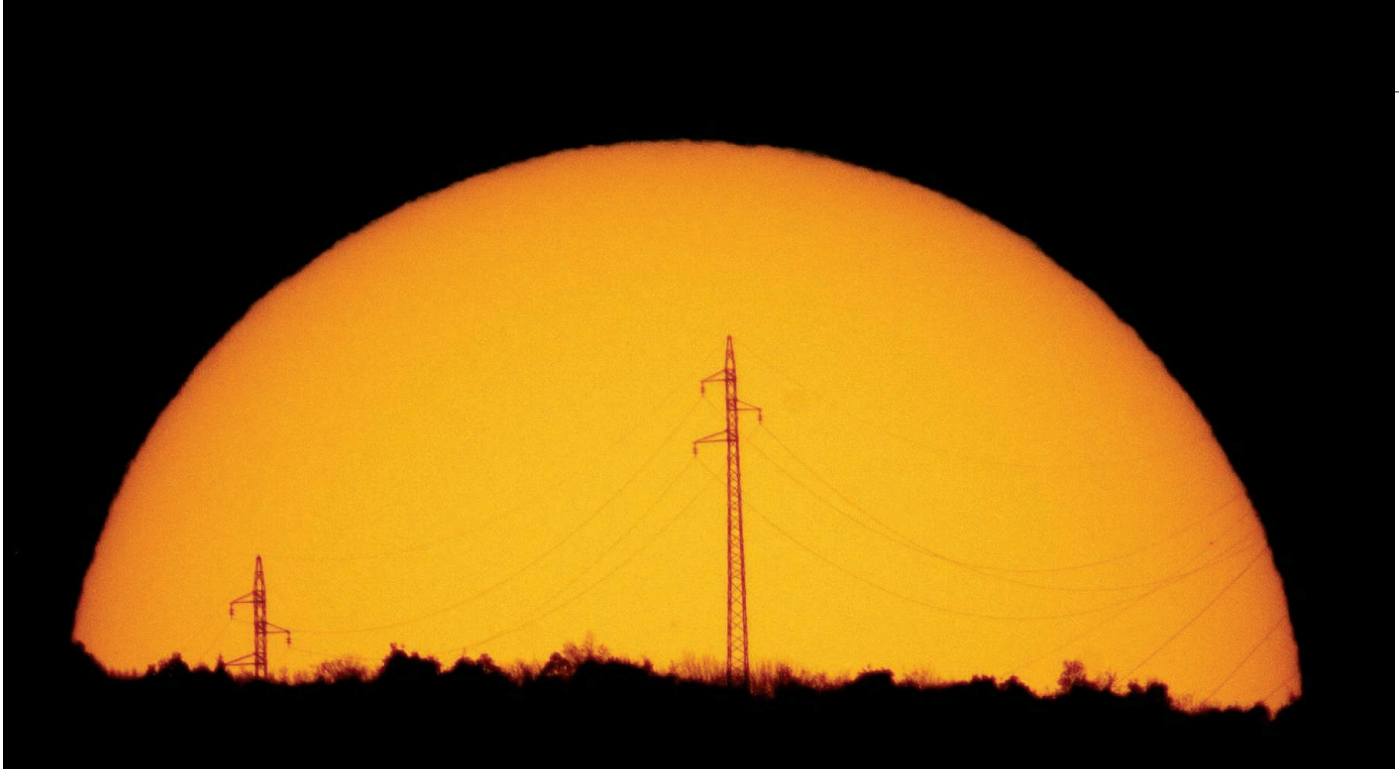
**ONE THREAT TO CIVILIZATION** could come not from too little sun, as in Leiber's story, but from too much. Bill Murtagh has seen how it might start. On the morning of 23 July 2012, he sat before a colorful array of screens at the National Oceanic and Atmospheric Administration's Space Weather Prediction Center in Boulder, Colorado, watching twin clouds of energetic particles—known as a coronal mass ejection (CME)—erupt from the sun and barrel into space. A mere 19 hours later, the solar buckshot blazed past the spot where Earth had been just days before. If it had hit us, scientists say, we might still be reeling.

Now the assistant director of space weather at the White House Office of Science and Technology Policy in Washington, D.C., Murtagh spends much of his time pondering solar eruptions. CMEs don't harm human beings directly, and their effects can be spectacular. By funneling charged particles into Earth's magnetic field, they can trigger geomagnetic storms that ignite dazzling auroral displays. But those storms can also induce dangerous electrical currents in long-distance power lines. The currents last only a few minutes, but they can take out electrical grids by destroying high-voltage transformers—particularly at high latitudes, where Earth's magnetic field lines converge as they arc toward the surface.

The worst CME event in recent history struck in 1989, frying a transformer in New Jersey and leaving 6 million people in Quebec province in Canada without power. The largest one on record—the Carrington Event of 1859, named after the U.K. astronomer who witnessed the accompanying solar flare—was up to 10 times more intense. It sent searing currents racing through telegraph cables, sparking fires and shocking operators, while the northern lights danced as far south as Cuba.

"It was awesome," says Patricia Reiff, a space physicist at Rice University in Houston, Texas. But if another storm that size struck today's infrastructure, she says, "there would be tremendous consequences."

Some researchers fear that another Carrington-like event could destroy tens to hundreds of transformers, plunging vast



Electrical surges due to a solar storm shocked telegraph operators in 1859; today, they could wreak havoc on power grids and electronics.

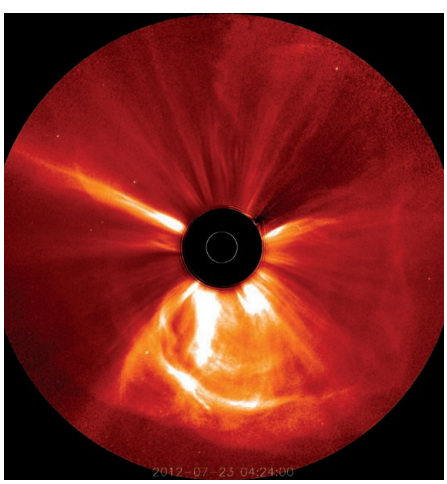
portions of entire continents into the dark for weeks or months—perhaps even years, Murtagh says. That's because the custom-built, house-sized replacement transformers can't be bought off the shelf. Transformer manufacturers maintain that such fears are overblown and that most equipment would survive. But Thomas Overbye, an electrical engineer at the University of Illinois, Urbana-Champaign, says nobody knows for sure. "We don't have a lot of data associated with large storms because they are very rare," he says.

What's clear is that widespread blackouts could be catastrophic, especially in countries that depend on highly developed electrical grids. "We've done a marvelous job creating a great vulnerability to this threat," Murtagh says. Information technologies, fuel pipelines, water pumps, ATMs, everything with a plug would be rendered useless. "That's going to affect our ability to govern the country," Murtagh says.

A major event could occur within our lifetimes. Research suggests that Carrington-like storms strike Earth once every few centuries; a recent study found a 12% chance that such a storm will occur in the next decade.

But at least we will see it coming. Solar telescopes spot CMEs right when they form, and spacecraft stationed a million miles from Earth measure critical parameters as they pass by. Armed with information like the orientation of a CME's magnetic field, scientists can tell whether the particle cloud will flow around Earth like "a rock in a river," Reiff says, or whether the field will connect with Earth's to stir up a geomagnetic storm. Forecasters can then issue alerts 30 minutes to an hour before the CME hits.

Such warnings are useful only if governments and grid operators are poised to respond, and countries around the world have just started to take the threat seriously. Last year, the White House released a comprehensive National Space Weather Strategy and an accompanying Action Plan laying out the need to reduce vulnerability and improve preparedness. A bipartisan bill to turn parts of the plan into reality will soon go before the Senate.




In 2012, satellites tracked this coronal mass injection from the sun as it barely missed Earth.

One pillar of the plan is to fortify the electric grid. Spurred by regulatory authorities, operators have already begun taking stock of vulnerable components and critical assets. The next step will be to protect the grid by installing current-blocking devices such as series capacitors, already common in the western United States because they aid long-distance power transmission, and by developing emergency procedures for manipulating power loads to limit transformer damage. Overbye says the power industry's swift response has been encouraging.

But full protection against a Carrington-like event might never be feasible, Overbye says, simply because of the cost. Instead, operators may react to an impending mega-storm by preemptively shutting down large portions of the grid to save transformers, embracing short-term devastation to avert a long-term disaster.

**FOR ANOTHER MENACE FROM THE SKY**—an impact by a large asteroid or comet—there is no way to limit the damage. The only way for humanity to protect itself, researchers say, is to prevent the collision altogether.

"That's something that we as a species can absolutely never, ever, ever let happen," Ed Lu says. "That's the end of human beings." In 2002, Lu, a former astronaut, founded the B612 Foundation in Mill Valley, California—a private organization that works to protect the planet from near-Earth objects, or NEOs.

Everyone knows about the 10-kilometer-wide asteroid that helped destroy the dinosaurs, but even something a fraction of that size could devastate humanity, says Michael Rampino, an earth scientist at New York University in New York City. The impact site would be obliterated, and massive earthquakes and tsunamis could radiate across the planet. But the lingering effects would prove most devastating. Models suggest that, depending on the speed and angle of approach, an object as small as 1 kilometer

PHOTOS: (TOP TO BOTTOM) NASA/MARTIN STOJANOVSKI; NASA

wide could throw up enough pulverized rock to block out the sun for months. Adding to the pall would be soot from wildfires ignited by debris falling back to Earth. "All this stuff coming back into the atmosphere heats up, and it's like setting your oven on broil," Rampino explains. Together, the smoke and dust would cast the planet into a so-called impact winter, causing crop failures and mass starvation.

Fortunately, asteroids of this size strike Earth only about once every few million years, and "dino killers" only once every 100 million years or so. Averaged annually, your chance of dying because of an impact is only slightly higher than that of perishing in a shark attack, says Mark Boslough, a physicist at Sandia National Laboratories in Albuquerque, New Mexico. But, like sharks, it only takes one to do the trick.

That's why, in 1998, NASA launched the Spaceguard survey at the request of Congress. The goal was to enlist astronomers to identify 90% of the estimated 900-plus NEOs bigger than 1 kilometer—a goal the agency officially met in 2010. Ongoing efforts now aim to find any remaining giants and tag 90% of bodies larger than 140 meters by 2020, although NASA says it won't meet the deadline. Of the nearly 15,000 NEOs discovered so far, none are currently on a collision course with Earth. Eventually, however, an Earth-bound NEO of some size will confront humanity with a disaster movie scenario. And when that day comes, "it's going to go from science fiction to science real pretty rapidly," Lu says.

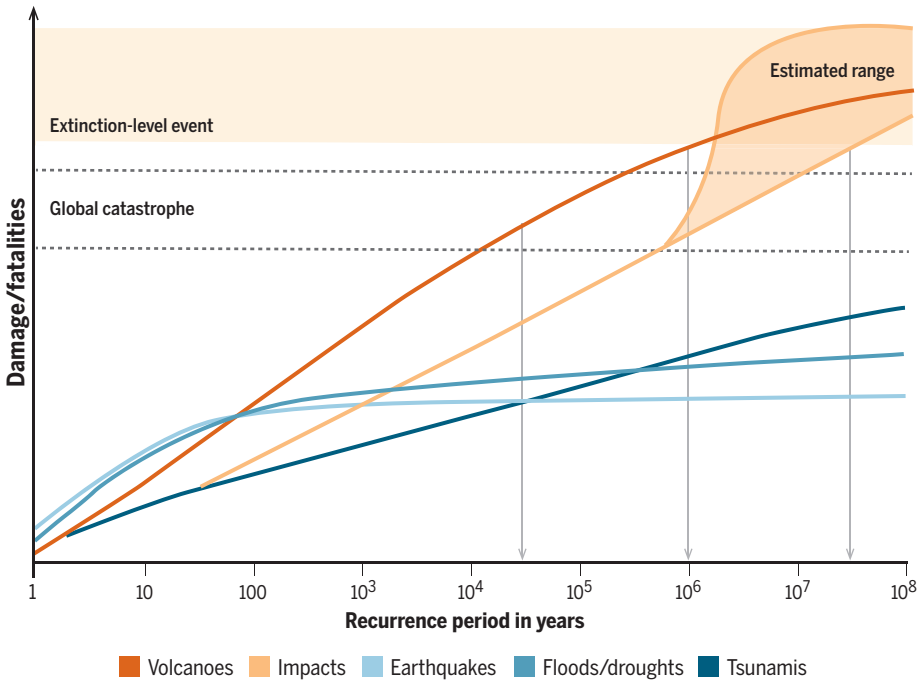
Science is already on the case. In *Defending Planet Earth: Near-Earth Object Surveys and Hazard Mitigation Strategies*, a 2010 report by the U.S. National Research Council, researchers highlighted several potential options for fending off an interloper, given a few decades of warning. We could whack it off course by ramming it with a spaceship or two, slowly alter its orbit with the sustained gravitational pull of a spacecraft called a gravity tractor, or blast it with nuclear explosions.

Right now, these planetary defense strategies exist mainly on paper, but some may see real-world tests in the next decade. NASA, the European Space Agency, and other partners are exploring a joint mission called AIDA (Asteroid Impact and Deflection Assessment) to test the impactor method on the asteroid Didymos when it passes near Earth in October 2022. NASA has also announced plans to use an enhanced gravity tractor—in which the spaceship collects material from the asteroid to increase its mass—on its Asteroid Redirect Mission, which was set to launch in 2021 but now faces funding setbacks. In the event of an actual threat, many researchers favor a combination of these techniques, just to be safe.

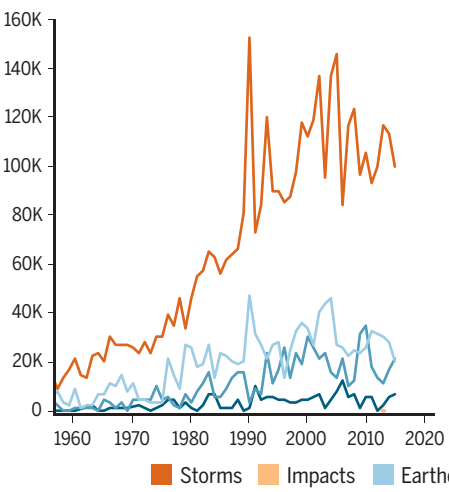
## Just a matter of time

Over thousands to millions of years, supervolcanoes and extraterrestrial impactors pose the greatest threats to humanity (top graph). Yet researchers and planners focus almost exclusively on more common hazards with only moderately severe conseqences (second and third graphs).

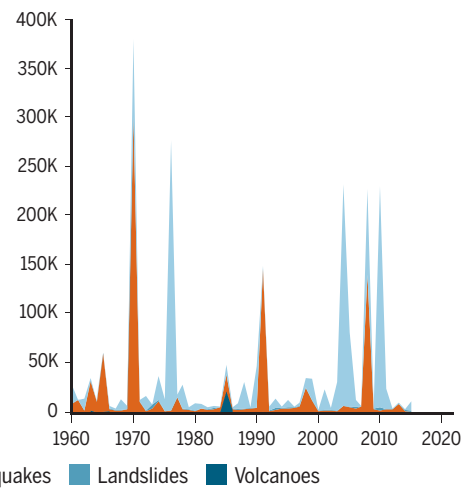
### Relative severity over time




### Number of recorded disasters



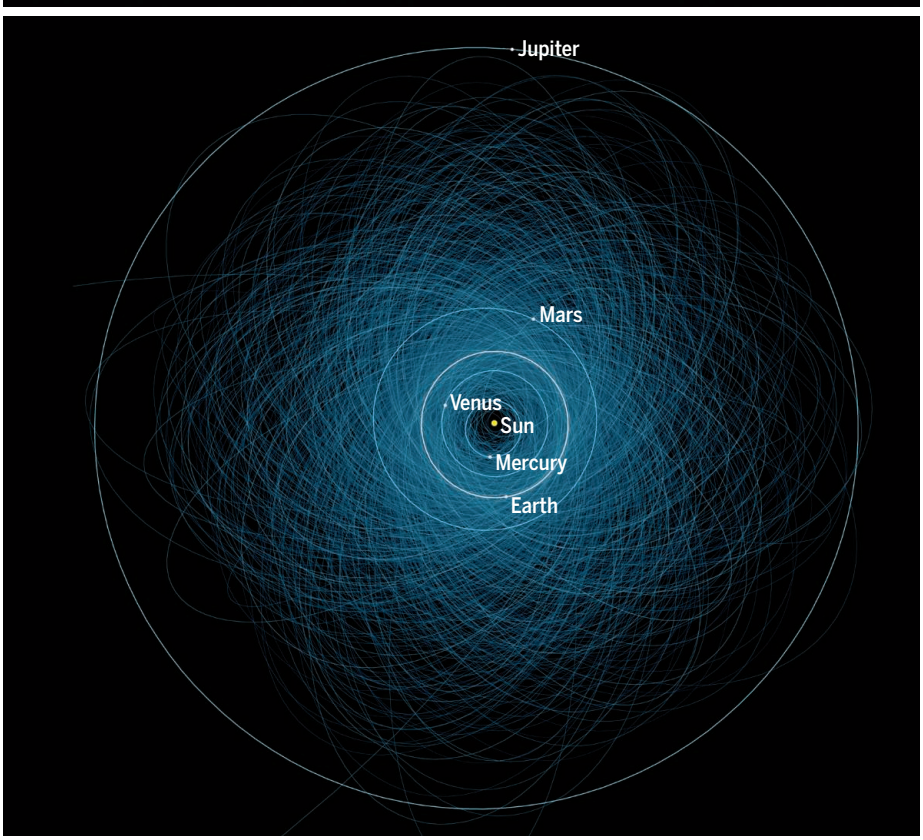

### Estimated deaths




But for objects larger than 1 kilometer across—and for comets, which could appear with little notice—some scientists think the nuclear option is the only option. The idea would be to jolt the body, not blow it up, which could do more harm than good. Although the 1967 United Nations Outer Space Treaty currently bars sending nuclear weapons into space, scientists already have a good understanding of the technology, and last year, NASA and the Department of Energy announced a joint effort to hone its use against asteroids. Ultimately, NASA's Planetary Defense Coordination Office, established earlier this year, will decide when and how the United States should respond to a potential impact.

**THE MOST INEXORABLE THREAT** to our modern civilization, however, is homegrown—and it strikes much more often than big cosmic impacts do. Every 100,000 years or

CREDITS: (GRAPHIC) G. GRULLÓN/*SCIENCE*; (DATA SOURCES) TOP GRAPH: PLAG *ET AL. EXTREME GEOHAZARDS* (2014) © ESF; NATURAL DISASTERS GRAPH: EM-DAT: THE INTERNATIONAL DISASTER DATABASE; FATALITIES GRAPH: EM-DAT: THE INTERNATIONAL DISASTER DATABASE




The Pan-STARRS telescope (top) on Maui in Hawaii is part of an astronomical network that scans the night sky for bodies that could someday collide with Earth. So far, astronomers have spotted almost 15,000 objects in Earth's neighborhood (bottom), including hundreds more than 1 kilometer across.

so, somewhere on Earth, a caldera up to 50 kilometers in diameter collapses and violently expels heaps of accumulated magma. The resulting supervolcano is both unstoppable and ferociously destructive. One such monster, the massive eruption of Mount Toba in Indonesia 74,000 years ago, may have wiped out most humans on Earth, causing a genetic bottleneck still apparent in our DNA—although the idea is controversial.

By geological convention, a supervolcano is one that produces an explosive eruption of more than 450 cubic kilometers of magma—roughly 50 times more than the eruption of Indonesia's Mount Tambora in 1815, and 500 times more than the Philippines' Mount Pinatubo in 1991. Geologists read the histories of such blasts in deposits of erupted material called tuff, and the rock record shows that supervolcanoes tend to be repeat offenders. Locations that remain active today include Toba, the Yellowstone hot spot in the northwestern United States, the Long Valley Caldera in eastern California, the Taupo Volcanic Zone in New Zealand, and several spots in the Andes.

None of these danger zones now poses a threat. But in the event of another eruption, everything within a hundred kilometers or so would be incinerated, and ash would blanket continents. Just a few millimeters of the stuff can kill crops; a meter or more can make land unusable for decades, says Susanna Jenkins, a volcanologist at the University of Bristol in the United Kingdom. Ash can also crush buildings, foul water supplies, clog electronics, ground airplanes, and irritate lungs.

These regional impacts could ripple around the world in unexpected ways. Even the minor disruption in air traffic following the 2010 eruption of Iceland's Eyjafjallajökull—a far cry from a supervolcano—caused millions of dollars in losses for Kenyan farmers, whose perishable exports to Europe went to waste. "The more interconnected our society becomes, the more vulnerable we are to something even quite small that happens over on the other side of the world," says Hazel Rymer, a volcanologist at The Open University in Milton Keynes, U.K.

Most far-reaching of all, however, would be the effects on global climate, which would resemble those of a large asteroid impact. Sulfate aerosols injected into the stratosphere by a supereruption could drop temperatures over much of Earth by 5°C to 10°C for up to a decade, devastating global agriculture.

Just how bad things would be is hard to say. "Volcano science is based on experi-

ence," says Ben Kennedy, a volcanologist at the University of Canterbury in Christchurch, New Zealand, and scientists have never witnessed a supervolcano. Knowledge of smaller eruptions can help, but it may be an unreliable guide. For instance, although supereruptions probably produce loads of sulfate aerosols, the aerosols may be larger and rain out faster than those from smaller eruptions, according to research by Claudia Timmreck, a climate modeler at the Max Planck Institute for Meteorology in Hamburg, Germany, and others. Timmreck's team has also found that for midlatitude volcanoes like Yellowstone, the season in which the eruption occurs determines how widely its aerosols spread.

The biggest uncertainties surround potential warning signs. Researchers think that widespread clues such as earthquakes, increased outgassing, and ground deformation due to rising magma would precede a major eruption. This unrest would begin months, if not many years, in advance, theoretically affording ample time to evacuate residents and set emergency response plans in motion. However, scientists would struggle to decide when to sound the alarm, says Jacob Lowenstern of the U.S. Geological Survey in Menlo Park, California, the scientist-in-charge of the Yellowstone Volcano Observatory. "It's going to be hard for scientists to convince themselves just because of our only partial understanding of the complexity of the processes that are taking place," he says.

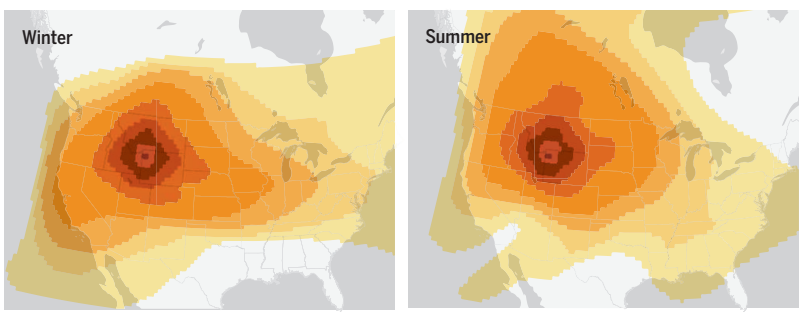
Then there are the political challenges of responding to the threat. The 1985 eruption of Nevado del Ruiz in Colombia killed 23,000 people, in part because the government ignored scientists' forecasts. False alarms can cause trouble, too. In the 1980s, geologic unrest caused officials to warn that California's Long Valley Caldera could erupt. It didn't, but local real estate values tanked and the economy suffered.

The challenge for scientists is to tell which indicators portend a catastrophic eruption instead of a small one—or none at all. "We're terribly good at recognizing precursors after the event," Rymer says. For now, researchers say, their best bet is to continue studying the plumbing that feeds volcanoes and to squeeze as much information as possible from smaller future eruptions before the next supervolcano lets loose.

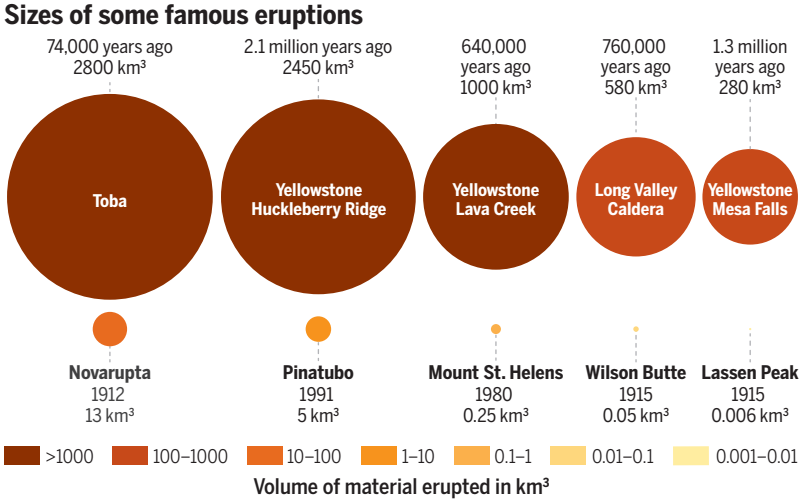
## Supervolcanoes, the homegrown menace

Volcanic blasts of the past dwarf any in recent history. One still-active U.S. site, Yellowstone, will blanket much of North America with ash if another supereruption ever explodes there.

### Modeled ash clouds from a Yellowstone supereruption, by season




### Sizes of some famous eruptions




CREDITS: (GRAPHIC) G. GRULLÓN/*SCIENCE*; (DATA SOURCES) YELLOWSTONE ERUPTION GRAPHIC: LARRY G. MASTIN; ERUPTIONS GRAPHIC: USGS

**IN THE END**, no amount of research can do much to prevent or mitigate supervolcanoes, or other freak events such as nearby supernova explosions and cosmic blasts of gamma rays. Our only hope of surviving them is a fallback plan. And the bottom line in that plan is food.

At least two scientists have already sketched out a blueprint. In their 2015 book *Feeding Everyone No Matter What*, David Denkenberger and Joshua Pearce propose several ways to feed billions of people without the help of the sun.

Denkenberger, an architectural engineer at Tennessee State University in Nashville, started moonlighting as a catastrophe researcher a few years ago after reading that fungi may have thrived after previous mass extinctions. If humans ever face a similar threat, he thought, "Why don't we just eat the mushrooms and not go extinct?"

Indeed, people could grow mushrooms on leaf litter and on the trunks of trees killed by the disaster, Denkenberger says. Even better would be raising methane-digesting bacteria on diets of natural gas, or converting the cellulose in plant biomass to sugar, a process already used to make biofuel. Denkenberger and Pearce—an engineering professor at Michigan Technological University in Houghton—calculate that by retrofitting existing industrial plants, survivors of the catastrophe could produce enough of such alternative foods to feed the world's population several times over.

Of course, a few other ingredients would have to survive as well: infrastructure, international cooperation, and the rule of law. Whether human society endures or snaps is the unknown on which everything else could hinge, says Seth Baum, executive director of the Global Catastrophic Risk Institute in New York City, a nonprofit think tank whose researchers include Denkenberger.

"How would we fare? I think the only reasonable answer one can give to the question at this time is that we have absolutely no idea," Baum says. To him, social resilience after a catastrophe is just another question for scientists to address, instead of leaving it to dystopian writers and doomsday preppers.

Not that he has anything against survivalists. "As much as they might seem silly on television, I'm actually a little happier knowing that there are people out there doing that stuff," Baum says. He quickly adds, "Hopefully it'll never come down to just that." ■

*Julia Rosen is a freelance writer in Portland, Oregon.*




# DOOMSDAY MACHINES

## Disaster simulators that whoosh, gush, and rumble can do things supercomputers (so far) can't match

*By* **Warren Cornwall**, *in San Diego, California*


The "Wall of Wind" at Florida International University in Miami can blow as fiercely as a category-5 hurricane.

PHOTO: ROBERT SULLIVAN

For the past year, Tara Hutchinson has been trying to figure out what will happen to a tall building made from thin steel beams when "the big one" hits.

To do that, she has erected a six-story tower that rises like a lime-green finger from atop a shrub-covered hill on the outskirts of San Diego, California. Hundreds of strain gauges and accelerometers fill the building, so sensitive they can detect wind gusts pressing against the walls. Now, Hutchinson just needs an earthquake.

In most of the world, this would be a problem. Even here, where a major fault runs right through downtown, the last quake of any note struck 6 years ago and was centered in nearby Mexico. But Hutchinson, a structural engineering professor at the University of California (UC), San Diego, doesn't need plate tectonics to cooperate. This summer she has an appointment at one of the world's biggest earthquake machines.

This device—a sort of bull ride for buildings—is one in a network built around the United States over the past 15 years to advance natural disaster science with more realistic and sophisticated tests. Costing more than $280 million, the National Science Foundation (NSF) initiative has enabled scientists to better imitate some of the most powerful and destructive forces on Earth, including earthquakes, tsunamis, and landslides.

The work has led to new building standards and better ways to build or retrofit everything from wharves to older concrete buildings. Scientists have gained insights into how quakes damage pipes in walls and ceilings and how to help quake-proof highway ramps, tall steel buildings, parking garages, wooden homes, and brick walls, to name a few.

That expansion continues today. In a new $62 million, 5-year program, the network of doomsday machines is expanding to simulate hurricanes and tornadoes and is joining forces with computer modeling to study how things too big for a physical test, such as nuclear reactors or an entire city, will weather what Mother Nature throws at them.

**CREDIT CALIFORNIA'S** Northridge earthquake for helping set this in motion. The 1994 quake, centered near Los Angeles, killed 72 and cost an estimated $25 billion in damages. In its aftermath, a report commissioned by Congress warned that the country needed a more systematic approach to studying how to reduce damage from earthquakes. NSF responded with the $82 million Network for Earthquake Engineering Simulation. The money funded a construction spree at 14 sites around the country. Another $200 million paid for operating the sites through 2014. That included UC San Diego, which unveiled the world's largest outdoor shake table in 2004.

Descriptions of these disaster labs are often couched in superlatives: the biggest, the longest, the most powerful. In addition to the San Diego facility, the projects funded under the original program and its successor, the Natural Hazards Engineering Research Infrastructure (NHERI), include North America's largest wave flume for studying tsunamis at Oregon State University, Corvallis; the world's largest university-based hurricane simulator at Florida International University in Miami; and, at UC Davis, the world's biggest centrifuge for making scale models mimic the stresses on tons of buildings, rock, and dirt—crucial information for assessing how structures will weather earthquakes and landslides.

More than bragging rights is at stake. When it comes to learning how buildings cope with the forces generated in a natural disaster, size often does matter. For example, the way soil particles stick together, an important factor in landslide risks, depends on how much mass is pushing down on them. Similarly, it's nearly impossible to build accurate, tiny versions of rebar: steel rods embedded in concrete structures that are critical to building performance. Similar difficulties arise with measuring how hurricane-force winds interact with a building.

"You can't take a real building and scale it down to one-tenth and put it in a wind tunnel. The physics doesn't work," says Forrest Masters, a wind engineer at the University of Florida in Gainesville who directs his university's share of NHERI. That includes a machine capable of subjecting 5-meter-tall walls to the air pressures found in a 320-kilometer-per-hour hurricane, and a wind tunnel whose floor can be modified to see how different terrain influences the way wind interacts with structures.

Computer models too can fall short in accurately reproducing all the forces at play as, say, a bridge twists and sways in an earthquake. So many different pieces in the bridge are pulled in so many directions at once that it can fail in unpredictable ways, causing models to misrepresent reality. In 2010, a contest at the San Diego shake table pitted 41 teams of experts running models against a real-life test of a 7-meter-tall bridge column topped with 236 metric tons of concrete blocks. The computer results were all over the place, says Stephen Mahin, a structural engineer at UC Berkeley who helped orchestrate the event. On average, they underestimated how much the column would sway by 25%. "You can't quite trust the computer results yet," Mahin says.

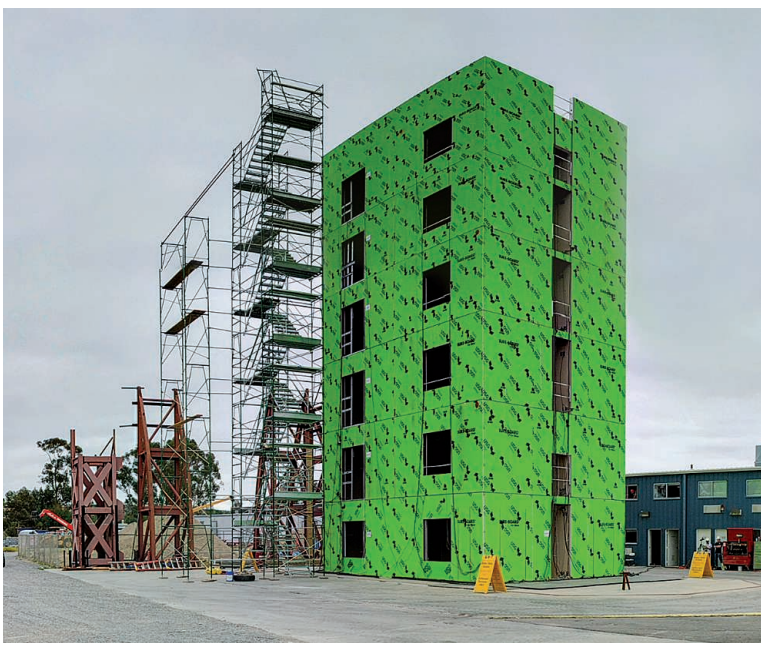
One morning in mid-May, Hutchinson inspects her building in the final stages of preparation for the test. She points to tiny gaps that have sprung open where metal ceiling joists meet the wall in a first-floor room. That happened during a minor, preliminary shake her team delivered to the building a day earlier. It's the kind of thing that could make a difference in how load is shared between pieces of the building, and how much damage the building suffers in the next temblor. And it wouldn't show up in a computer model.

"You're not going to account for every screw," she says. "Look at how subtle this damage is."

**DEVISING A MACHINE** that can pack the same wallop as a magnitude-8.0 earthquake or a category-5 hurricane isn't easy, or cheap. A look under the hood of San Diego's shake table illustrates the kind of mechanical muscle needed. Joel Conte, an engineering professor who oversees the shake table operations, leads the way into a cavernous underground room filled with machinery. A 20,000-liter metal tank holds the hydraulic fluid that drives the entire system. Two pumps slurp the fluid from there into a bank of 50 slender black cylinders reminiscent of street light poles at pressures reaching 34,000 kilopascals (5000 pounds per square inch). That high pressure is crucial, generating enough force to swiftly move an entire building.

Conte turns down a passageway, tracing the path of the fluid through steel pipes 30 centimeters across, and into a room dominated by a mass of steel resembling the hull of a flat-bottomed boat. This is the epicenter. A metal plate 5 centimeters thick, 12 meters long, and nearly 8 meters wide sits overhead, bolted to the steel underbelly. At either end, an actuator that looks something like a car's shock absorber, but is as thick as a man's torso, extends from this structure to the concrete wall. When the commands come from computers in a nearby building, the actuators will jerk to life, the hydraulic fluid driving them back and forth. The plate, pushed and pulled between them, will slide across metal sheets polished mirror-smooth at speeds of up to 1.8 meters per second. Voilà! Instant quake.

"The real world, you cannot count on it," Conte says. "You cannot say, 'Oh, I'm going to sit and wait for the next earthquake in front of this big building, and I'm going to invest a lot in sensors.' You may have to wait 30, 40, 50 years. So you produce an earthquake."



A building awaits its ordeal on the shake table at the University of California, San Diego.

PHOTO: ERIK JEPSEN/UC SAN DIEGO

Since its construction for $10 million, the shake table has tested a four-story concrete parking garage, a wind turbine, and a five-story concrete building complete with elevator and stairs, among other things. The tests have shown that special inserts can increase resilience by allowing a building to move over its foundation and that modular concrete floors can behave erratically unless they have additional reinforcement. They have also revealed how tall, wood-framed buildings fail and how reinforcements can strengthen old brick buildings.

Back in his office, Conte gleefully clicks through the "best of" video highlights. A four-story wood building twists and splinters to the ground. A parking garage teeters back and forth like a rocking chair. A split screen shows two identical rooms filled with hospital beds and medical equipment. One is in a building outfitted with padded foundations that help it absorb an earthquake's shock; the other isn't. As the video runs, beds in the regular building suddenly lurch back and forth before toppling over. In the other, they barely move.

In the current test, Hutchinson wants to see how a building six stories tall made from lightweight steel performs during and after an earthquake. She thinks it could do well, partly because it's lighter than a concrete building of the same height, giving it less mass to generate damaging forces during a quake. Today, building codes allow this type of construction to be just shy of 20 meters tall. But the tallest building really put to the test was only two stories high.

The structure, modeled after an apartment building, is destined for a multistage torture test. Hutchinson and her colleagues will first put it through a simulation of several quakes, including Northridge and a 2010 magnitude-8.8 in Chile. Then they will set fires in parts of the building to see how it holds up in a blaze triggered by quake damage. Then they will shake the building again in a mock aftershock, hard enough that it might collapse.

The results aren't just of academic interest. Sponsors of the test include manufacturers of the steel construction parts, the insurance industry, and state government. "There's nothing like a full-scale test," says Richard McCarthy, executive director for the California Seismic Safety Commission in Sacramento, a government commission that advises policymakers. It contributed $100,000 to the event, he says, partly with an eye toward potential changes to building codes governing construction using these materials.

Conte is now lobbying state officials for a $14 million upgrade that would allow the machine to run even more realistic tests. Right now it can move only back and forth in two directions; new hardware would add up-and-down, side-to-side, and diagonal motions, enabling it to move in every direction—like the world's biggest shake table, an indoor facility in Miki, Japan.

Researchers at Oregon State University, Corvallis, unleash tsunamis in a wave basin.

**SCIENTISTS ARE TRYING** to go even bigger by marrying such physical tests with computer models. The resulting "hybrid" simulations can test massive structures too big to fit inside any test facility, says James Ricles, a civil engineer at Lehigh University in Bethlehem, Pennsylvania. His lab, which is part of the NSF network, tests well-understood parts of a structure with computer models but stages physical tests for parts that the models can't handle. In a feedback loop measured in milliseconds, sensors from the physical test send data to the model, which adapts and sends new signals that tell the machines driving the physical test how to tweak their next moves.

Ricles's lab simulated the behavior of an elevated highway during an earthquake by physically testing the concrete columns while testing a virtual model of the bridge deck in a computer. He recently applied the same strategy to testing a design meant to allow a steel building to rock back and forth rather than bend during a quake. A four-story chunk of the building stood in the lab; the rest of it existed only in the microprocessors of a computer.

Destruction is a definite part of the work's appeal, says Gilberto Mosqueda, an engineering professor who runs hybrid tests at UC San Diego: "You build these models, and essentially you shake them till you break them." But the mountains of data generated by the tests also open the way to more sophisticated numerical models that could one day do some of the work of the doomsday machines.

Whereas the earlier NSF program focused on big testing platforms, the NHERI initiative is putting more money into the virtual side. The University of Texas, Austin, won $13.7 million to build a data repository and software platform to store information from years of field tests. In the future, engineers should be able to tap data in the digital repository to boost the accuracy of their computer models. And NSF will soon issue an $11 million award for a computational modeling and simulation center.

"Will we get to the point where we can just model everything and have confidence? That may still be a long way off," says Joy Pauschke, a structural engineer and director of the NSF program that funds the testing work in Arlington, Virginia. "But hopefully as we test and improve models, we start moving towards having better capabilities with the computational modeling."

Berkeley's Mahin—whose 2010 contest exposed the shortcomings of models—now also foresees bright prospects for modeling. Advances in machine learning and cloud computing, he predicts, will lead to models capable of simulating not just single buildings but entire communities. Unleashing "virtual disasters" could then enable researchers and government officials to grasp the region-wide effects of a major quake or storm and decide which measures today would prevent the most damage.

"In 20 years, you can model a whole city in a very complicated way, I think," Mahin says. "There's a great hope this analysis can help mitigate the damage from future natural disasters." ■

PHOTO: © AURORA PHOTOS/ALAMY STOCK PHOTO

REVIEW

# Human influence on tropical cyclone intensity

Adam H. Sobel,[1,2]* Suzana J. Camargo,[2] Timothy M. Hall,[3] Chia-Ying Lee,[4] Michael K. Tippett,[1,5] Allison A. Wing[2]

**Recent assessments agree that tropical cyclone intensity should increase as the climate warms. Less agreement exists on the detection of recent historical trends in tropical cyclone intensity. We interpret future and recent historical trends by using the theory of potential intensity, which predicts the maximum intensity achievable by a tropical cyclone in a given local environment. Although greenhouse gas–driven warming increases potential intensity, climate model simulations suggest that aerosol cooling has largely canceled that effect over the historical record. Large natural variability complicates analysis of trends, as do poleward shifts in the latitude of maximum intensity. In the absence of strong reductions in greenhouse gas emissions, future greenhouse gas forcing of potential intensity will increasingly dominate over aerosol forcing, leading to substantially larger increases in tropical cyclone intensities.**

Theory and numerical simulations suggest that human emissions of greenhouse gases, acting on their own, should have already caused a small increase in tropical cyclone (TC) intensities globally. The same theory and simulations indicate that we should not expect to be able to discern this increase in recent historical observations because of the confounding influences of aerosol forcing (which acts to oppose greenhouse gas forcing) and large natural variability (which compromises trend detection). Current expectations for the future are that aerosol forcing will remain level or decrease while greenhouse gas forcing continues to increase, leading to considerable increases in TC intensity as the climate warms further. Relatively low confidence in observed trends and greater confidence in projections of future trends are both consistent with our current understanding.

We first review projections of future TC activity, considering not only intensity but other measures including TC frequency (the number of storms per year), TC-induced precipitation, and coastal flooding. Future projections underpin any discussion of human influence in the present. The influence of greenhouse gas emissions is expected to be greater in the future, and the literature regarding the mechanisms of that influence focuses more on future scenarios. We then discuss recent historical trends in both TC activity and potential intensity (PI), an environmental parameter that encapsulates much of our theoretical understanding of the control that climate exerts on TC intensity.

[1]Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027, USA. [2]Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY 10964, USA. [3]NASA Goddard Institute for Space Studies, New York, NY 10025, USA. [4]International Research Institute for Climate and Society, Columbia University, Palisades, NY 10964, USA. [5]Center of Excellence for Climate Research, Department of Meteorology, King Abdulaziz University, Jeddah, Saudi Arabia.
*Corresponding author. Email: ahs129@columbia.edu

## Projected influence of climate change on tropical cyclones from models and theory

After an initial period of disagreement [e.g., (*1*–*5*)], more recent assessments and reviews are in broad agreement regarding future projections of various aspects of TC activity and the confidence associated with each aspect (*6*–*9*). An important source of new results informing the present consensus has been high-resolution global models, which simulate TCs with greater fidelity than did earlier models (*10*). A parallel development has been a much more nuanced understanding, based on multiple lines of evidence, of the relationship between TCs and their local environment.

PI is a useful parameter for understanding TC intensity as a function of the large-scale environment. It has multiple formulations that differ in detail but are broadly similar; here we use the formulation of Bister and Emanuel (*11*). The PI is a function of both the sea surface temperature (SST) and the vertical profiles of temperature and humidity in the atmosphere above. The PI is derived from an explicit physical theory that predicts the maximum intensity that a TC can achieve, given those environmental conditions. Although the validity of PI theory has been critiqued on multiple grounds (*12*, *13*), it nonetheless appears, under detailed scrutiny, to hold up well enough to be useful (*14*), and it provides a strong theoretical underpinning for discussions of how tropical cyclones may change with climate.

In the simulations from both the Third (*15*, *16*) and the Fifth (*17*) Coupled Model Intercomparison Projects (CMIP3 and CMIP5), climate models predict robustly that PI should increase in future warming scenarios in most regions where tropical cyclone activity occurs. The eastern part of the North Atlantic, where PI is projected to decrease, is a notable exception. PI essentially measures the degree of disequilibrium between the ocean surface and the atmosphere; a warmer ocean or a cooler atmosphere (in an appropriately generalized sense, accounting for temperature, humidity, and vertical structure from the surface to the lower stratosphere) leads to greater PI. Both models and physical arguments (*16*, *18*, *19*) predict that these differences should increase with climate warming (Fig. 1). The largest predicted increases are on the margins of the tropics, particularly in the Atlantic and Pacific.

Although PI is a prediction only of the maximum intensity that a tropical cyclone can achieve in a given environment, it is expected to provide a useful guide to the statistical distribution of actual intensities achieved by real TCs. Most TCs do not achieve their PI because of a variety of negative influences not fully accounted for by the theory, such as vertical wind shear and entrainment of dry air into the storm [e.g., (*20*)] or ocean coupling [e.g., (*21*–*23*)]. Nonetheless, observational evidence (*24*–*26*) and high-resolution numerical simulations (*14*, *27*) suggest that changes in PI are associated with changes in the average intensities that storms achieve. We thus expect TC intensities to increase with warming, both on average and at the high end of the scale, so that the strongest future storms will exceed the strength of any in the past. Although factors such as wind shear may change so as to modulate the influence of PI in some regions [e.g., (*28*)], we expect these influences to be more variable globally, and we assume as a starting point that they will not systematically counteract the global effect of increasing PI. This expectation is consistent, at least qualitatively, with the results of the most convincing numerical simulations, which show TC intensities increasing in a warming climate (*29*–*32*).

The frequency of TC occurrence is much less well understood than is TC intensity. About 90 ± 8 TCs form on Earth each year (*33*, *34*). In contrast to the situation for intensity, no physical theory predicts this number, even to an order of magnitude, despite intense research activity on mechanisms controlling the genesis of individual TCs [e.g., (*35*, *36*)]. Subseasonal to interannual variations in TC frequency are widely diagnosed using semi-empirical genesis indices (*37*–*40*), but using these to predict future global changes requires a problematic out-of-sample extrapolation (*41*, *42*).

Global high-resolution models have tended to simulate reductions in TC frequency under a warming climate (*30*, *31*, *43*, *44*); the most likely explanations for this behavior at present (*38*, *42*) involve an increasing saturation deficit (the difference between actual and saturation specific humidity) as temperature increases and relative humidity stays approximately constant [e.g., (*45*)]. These arguments have not been adequately developed or tested, however, and some credible models predict increases in TC frequency with warming (*23*, *46*). Altogether, the projection that TC frequency will decrease with warming is considerably more uncertain than the projection that TC intensity will increase.

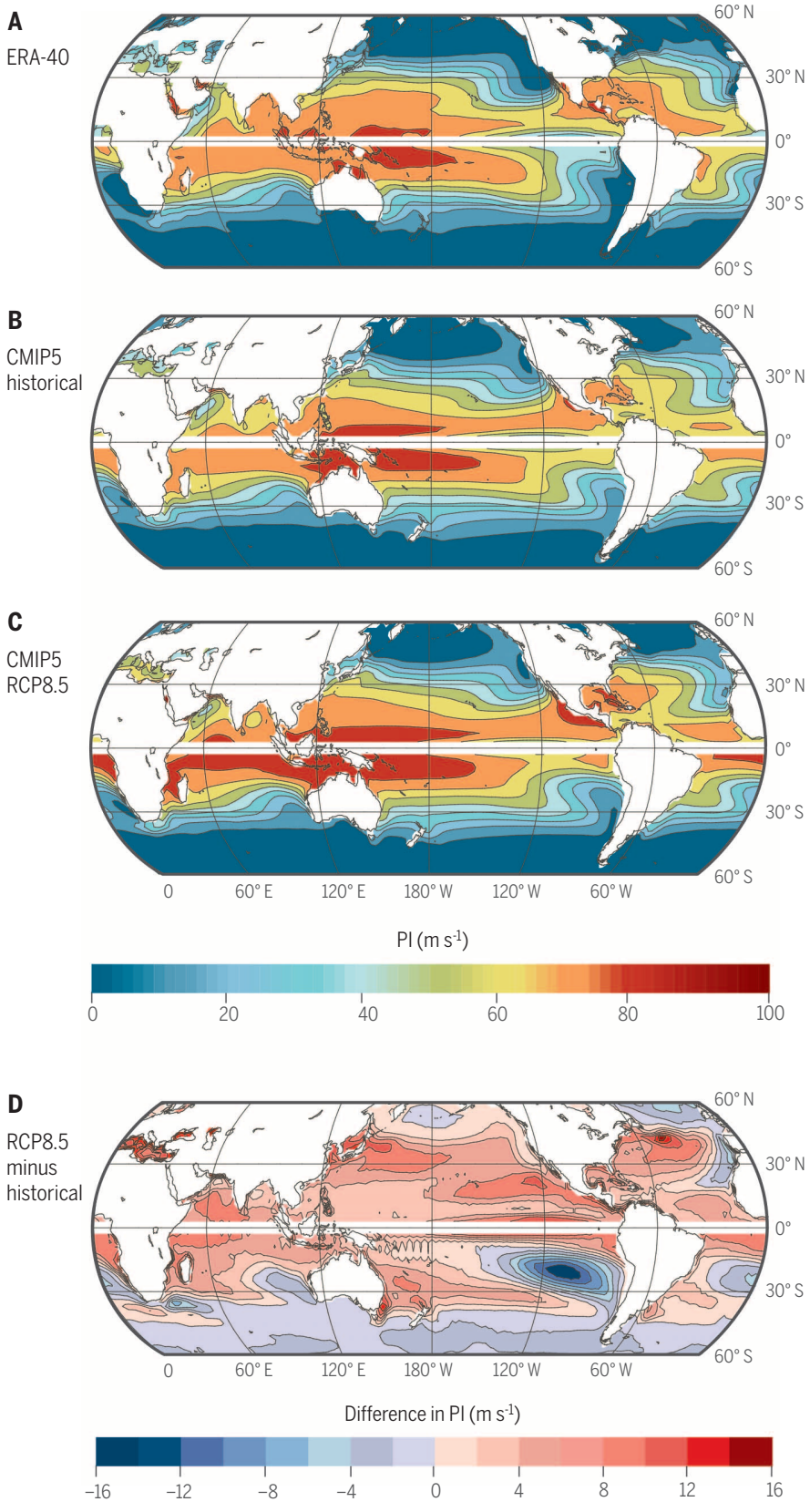



**Fig. 1. PI climatology.** (**A**) ERA-40 reanalysis (*84*) for 1971–2000, (**B**) CMIP5 historical runs (multimodel mean) for 1971–2000, (**C**) CMIP5 RCP8.5 projection (multimodel mean) for 2071–2100, and (**D**) values in (C) minus those in (B). The Northern (Southern) Hemisphere season is June to November (December to May). [Inspired by figure 1 in (*72*) and figure 4 in (*19*)]

Some other aspects of TC change associated with climate warming are relatively well understood. Increasing precipitation produced by TCs due to increasing water vapor content is expected with considerable confidence as the climate warms (*6*, *47*, *48*–*50*). Coastal flooding due to TC-induced storm surge is essentially certain to increase with warming as a consequence of sea level rise (*8*, *9*, *51*,), possibly compounded by increases in storm intensity or frequency [e.g., (*52*–*54*)].

## Observed trends

The detection of long-term trends in TC activity has been a subject of considerable debate. The validity of data from the earlier periods in the longest-term observational data sets has been strongly questioned [e.g., (*55*–*58*)], making any trends computed from records that include those periods disputable. Large natural variability, including substantial components with decadal and longer frequencies, further confounds trend detection in records, which in many cases are only a few decades long. These difficulties have led to findings of low confidence in observed TC trends in consensus assessment reports (*7*, *59*).

Perhaps the most persistent and provocative [though not unchallenged (*58*)] findings are that intensity increased in the past few decades at the upper end of the observed range, implying an increasing frequency of storms in categories 4 and 5 on the commonly used Saffir-Simpson scale (*60*–*62*), and that overall activity increased in the North Atlantic over a period of roughly three decades, beginning in the 1970s (*63*, *64*).

The North Atlantic changes are of great regional interest in North America. Their causes and implications for the future are the subject of debate. The increase in TC activity in the late 20th century coincided with both absolute and relative warming of the North Atlantic SST. The absolute warming is likely to continue, whereas the relative warming [which has a stronger influence on PI; for example, (*65*–*67*)] is not, at least not at the same pace. This implies that future increases, if they occur at all, will be more gradual. More in-depth arguments focus on the specific reasons for the Atlantic warming and associated PI changes, with roles for radiative forcing from greenhouse gases, aerosols, and ozone depletion, as well as internal variability in the Atlantic basin (*19*, *67*–*72*).

Recent analyses of trends in the most intense storms continue to show evidence of increases in their numbers, but they do not fully resolve the difficulties posed by the combination of data inhomogeneities and large natural variability. Kossin *et al.* (*62*) examined trends in the quantiles of lifetime maximum intensity (LMI) in two different data sets over the period 1982–2009, considering only storms whose LMI was 65 knots or greater (Fig. 2). Over this period, the global LMI distribution assessed from best-track data had positive trends in the mean (2 m $s^{-1}$ $decade^{-1}$) and in its quantiles (>2 m $s^{-1}$ $decade^{-1}$ for quantiles > 0.4). In contrast, the LMI distribution from a novel satellite-based data set, designed to be

temporally homogeneous, had a mean trend (<1 m $s^{-1}$ $decade^{-1}$) and quantile trends that were insignificant. A discontinuity correction further reduced the trends in the LMI distribution. Holland and Bruyère (*73*) performed an apparently similar analysis to that of (*62*) but obtained somewhat different results. They used a climate model–based measure of the anthropogenic influence on global mean temperature, rather than time, as a covariate. They focused on the period 1975–2010, during which that covariate appears to vary close to linearly with time, and found that the proportion of storms reaching categories 4 and 5 increased with the covariate. They obtained this result both when using the best-track data and the satellite-based intensity data set used in (*62*). Their result thus appears to be inconsistent with the findings of (*62*), unless their covariate differs substantially from a linear trend, which does not appear to be the case.

Should we expect increases in the frequency of the most intense storms to have occurred by now as a consequence of human-induced global warming? Simple considerations suggest that, given a PI increase, actual intensity changes should be most apparent in the highest quantiles of the intensity distribution (*74*). TC maximum intensities, when not limited by rapid decreases in PI (such as occurs at landfall, for example), are observed to be uniformly distributed between tropical storm intensity and hurricane intensity and uniformly distributed again (but with a different value of the probability) between hurricane intensity and the local PI (*24*, *25*). If the entire PI distribution were to be simply shifted to higher values, the upper bound on TC intensity would increase, whereas the lower bound in typical analyses—the wind speed threshold for either tropical storm or category 1 hurricane intensity, for example—would stay constant. This one-sided expansion of the range in TC intensities would imply larger changes in higher quantiles. This appears to be qualitatively consistent with quantile regression analyses that show significant increases (with time and SST) in the highest quantiles but little or no increase in lower quantiles (*61*, *62*).

Interpretation of any observed intensity increases along these lines is confounded, however, by uncertainties in PI trends. Robustly detectable trends in basin-average PI are found only in the North Atlantic, where both surface warming and, to some extent, tropical tropopause cooling have contributed to an increase in PI between 1980 and 2013 (*70*, *75*). Elsewhere, there are few statistically significant trends and large disagreements between data sets (*75*–*77*), reflecting the uncertainties in tropospheric temperature trends and the limitations of reanalysis and radiosonde data.

## The anthropogenic signal at present

As described above, there remains some disagreement between studies regarding recent historical trends in TC intensity. Similar inconsistency in PI trends from different data sets (outside the Atlantic) weakens our expectation on physical grounds that intensities should have increased. Even if we were to draw the conclusion that our confidence in these historical trends should be low, however, it would be inappropriate to go on to conclude that there is no human influence on TCs at present. To draw that conclusion would be a type II statistical error, conflating absence of evidence with evidence of absence.

Models and theory indicate that human emissions of greenhouse gases should already have caused modest increases in PI on the global scale compared with what would have occurred in the absence of those emissions. Figure 3 shows trends in PI predicted by CMIP5 climate models. These results are from simulations of the historical period, using all forcings (natural and anthropogenic), greenhouse gases only, and aerosols only. Details of the simulations and the calculations can be found in (*72*), the authors of which performed similar analyses for the North Atlantic; these differ only in that we show results for the entire Northern and Southern Hemispheres.

In these simulations, greenhouse gas and aerosol forcings have approximately equal and opposite influences on PI, resulting in no significant trend, until the most recent period, when the aerosol forcing stops increasing (*78*) while the greenhouse gas forcing continues to increase. This is the case although the average aerosol forcing—which, including cloud-aerosol effects and comparing the end of the period with the beginning, is on the order of 1 $Wm^{-2}$ [e.g., (*79*, *80*)]—is only about half the greenhouse gas forcing, so that the net forcing is positive and results in significant SST trends, even as it has little effect on PI. We expect aerosols to influence surface temperature primarily by reducing the shortwave radiative energy flux, whereas greenhouse gases do so by increasing the longwave flux. Single-column model calculations [figure 2b in (*18*)] suggest that shortwave forcing is more effective than longwave forcing, by about a factor of 2, at changing PI per unit SST change. The close cancellation between the two influences on PI over most of the historical record in these simulations appears broadly consistent with our expectations, given the model inputs. [Whether these aerosol forcings are correct is another question; Stevens (*80*) argued that they may be overestimates. We do not consider this important question here.]

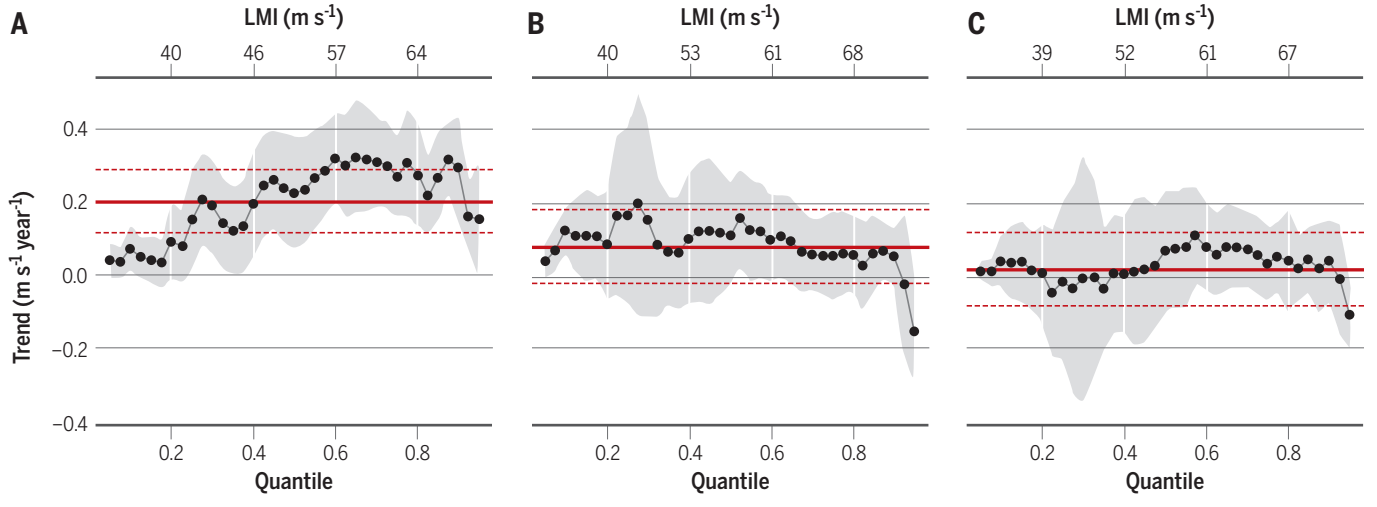



**Fig. 2. Global trends in the LMI quantiles of storms that achieved hurricane strength (LMI ≥ 33 m $s^{-1}$) in the period 1982–2009.** The left panel shows trends in the best-track data, the middle panel shows trends in the ADT-HURSAT (Advanced Dvorak Technique–Hurricane Satellite) record without an additional homogenization step to account for a discontinuity in the satellite data, and the right panel shows trends in the ADT-HURSAT record with the additional homogenization correction. The black dots represent the trends in the quantiles of the LMI distribution from 0.05 to 0.95 in steps of 0.025. Shading represents pointwise 95% confidence (two-tailed). The red solid line shows the (constant value) trend in the mean, as measured by ordinary least squares regression, and the red dashed lines show the confidence interval. The top axis shows the LMI values associated with the quantiles along the bottom axis. [From (*62*), used with permission (copyright American Meteorological Society)]

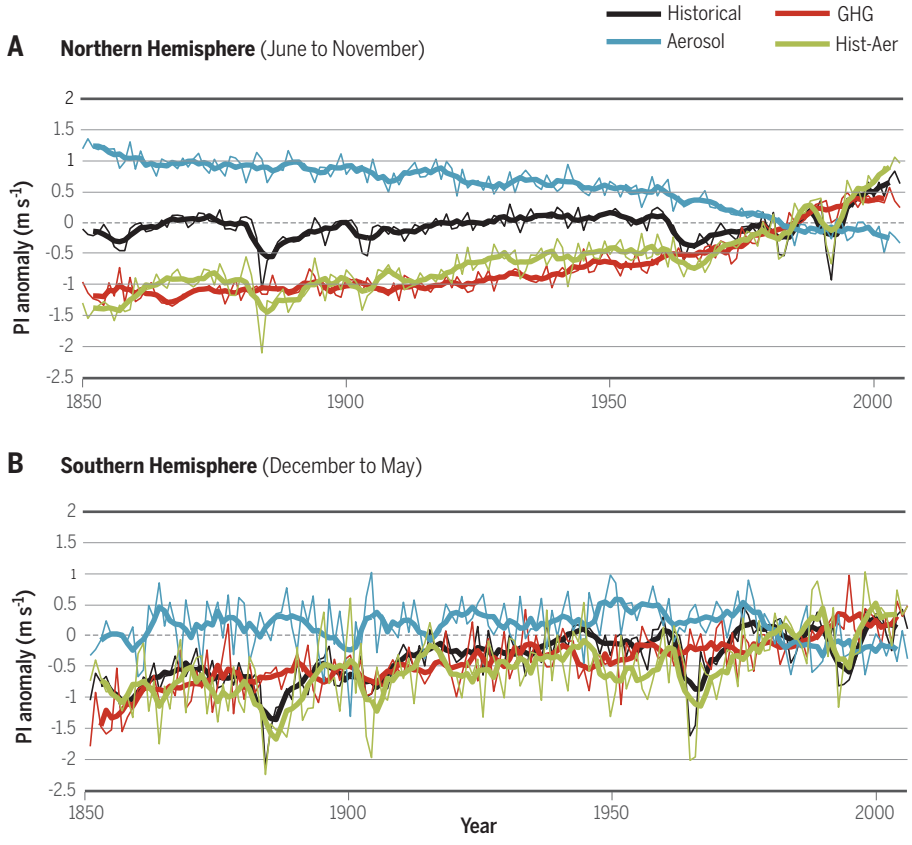



**Fig. 3. PI time series from CMIP5 historical runs with all forcings, greenhouse gas–only (GHG) runs, aerosol-only runs, and the difference between the historical and aerosol-only runs (Hist-Aer).** (**A**) Northern Hemisphere (0°N to 40°N). (**B**) Southern Hemisphere (0°S to 40°S). The thin lines show the multimodel mean for each season, and the thick lines show the 5-year running mean.

Aerosol lifetimes in the atmosphere are much shorter than those of the most important anthropogenic greenhouse gases, and future projection scenarios assume that greenhouse warming will exceed aerosol cooling in the future to an increasingly greater extent than in the past, unless very strong reductions in greenhouse gas emissions are implemented soon (*81*). We thus take the greenhouse gas–driven component of the PI perturbation in the present to be representative of the human influence as it will be increasingly manifest in the future.

When we ask whether there is a human influence on TC activity in the present, we should not stop at detection of trends with a confidence level of 95% (or any other specific level) against a null hypothesis of no trend. We might better ask whether the observations are consistent with less naive hypotheses based on models and theory. One such hypothesis might be based on PI. We might assume that on average, LMI is uniformly distributed between a lower bound and the PI, following (*24*). Kossin *et al.* (*62*) constructed a stochastic model based on this assumption; they showed that even in the presence of PI trends larger than those evident in observations and CMIP projections, trends in PI and the LMI distribution would be difficult to detect given the length of the reliable historical record, though trends in the most extreme part of the LMI distribution are more likely to be detected than trends in the middle of the distribution.

We might go even further and construct a similar null hypothesis for a bulk measure of TC activity such as the power dissipation index (PDI), defined as the cubed maximum wind speed of all storms integrated over their lifetimes and added together. If we neglect all variations in storm lifetime, storm frequency, and geographic storm-track distribution, we can construct a simple proxy, $\mathrm{PDI} = c\langle \mathrm{PI}^3 \rangle$, where the angle brackets represent basin averages and $c$ is a constant chosen to fit the recent historical average. [An accumulated cyclone energy (ACE) proxy would be the same, except with PI squared rather than cubed.] We do not claim that this is a correct model, just that it is a better null hypothesis than zero trend. The neglect of projected storm frequency changes might be justified as a starting point, given the lower confidence in these changes compared with those in intensity. More sophisticated hypotheses could be based on results from simulations of TCs in high-resolution global models, for example, or could incorporate projections of TC frequency change.

The lack of consistently significant trends in PI (outside the North Atlantic), in either observations or numerical simulations of the historical period, leads us to expect no trend in PDI under our simple null hypothesis. The CMIP5 simulations indicate that the lack of a trend is due to cancellation of aerosol and greenhouse gas effects. Even if the aerosol effects were entirely absent, however, trend detection would be challenged by the large natural variability.

We show in Fig. 4 the observed PDI in the Northern Hemisphere for the period 1950–2014, superimposed on the simple PI-based proxy described above, which we calculated by using the same CMIP5 simulations as in Fig. 3 for the historical period and the representative concentration pathway 8.5 (RCP8.5) scenario (*81*) for the future. Using this simple proxy as a baseline measure of our expectations, we would not expect to detect significant trends in PDI over the historical period, considering both the small trend in the PI-based proxy for PDI and the large variability in observed PDI (the model-based estimates have little variability because they are both basin and ensemble averages). Even in the absence of aerosol forcing, the trends would have been small compared with the shorter-term variability. These small and difficult-to-detect trends are is entirely consistent, however, with greater projected increases in PDI over the course of the 21st century under RCP8.5. The future trend in the PI-based proxy for PDI, shown in Fig. 4, is not a prediction for PDI, because multiple factors other than PI influence PDI (or any other bulk measure of TC activity) and are not accounted for by our simple proxy measure. The figure simply makes the general point that small (and thus difficult-to-detect) changes in TC activity up to the

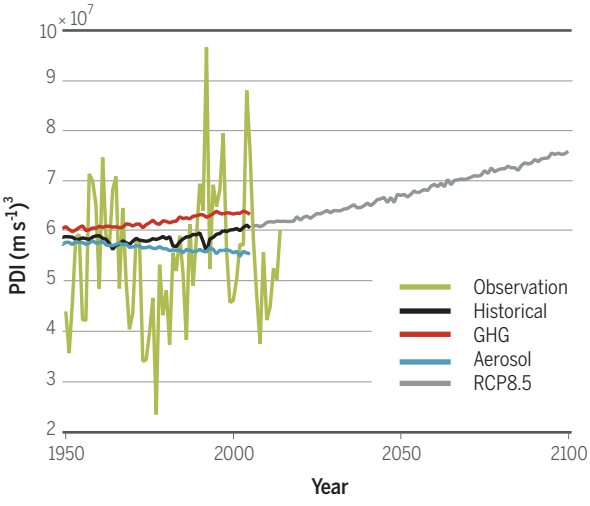



**Fig. 4. Observed and proxy PDI for the Northern Hemisphere.** The proxy was constructed from the basin- and ensemble-mean PI distribution, using CMIP5 simulations for the historical period (as in Fig. 3) and the RCP8.5 scenario for the future.

present and substantially larger changes in the future are both consistent with control of TC intensity (as measured by PI) by the large-scale environment.

Assuming that the average intensity of real TCs scales with basin-wide PI, as our simple proxy does, implies an assumption that TC tracks sample the PI distribution within the basin in the same way over time. Observed trends in the latitude of LMI have been detected, however, challenging this assumption and complicating the interpretation of both historical PI trends and future projections. The results of (*82*) indicate that in the recent historical past, TC tracks have shifted systematically poleward to regions of lower PI, to an extent that approximately compensates for the PI increases at fixed locations, so that the actual PI experienced by the storms has stayed approximately constant, despite basin-wide increases (*77*). It is not clear whether the poleward shift is a response to radiative forcing, but simulations for the western North Pacific suggest that it is, at least in part (*83*). The fact that projections show the largest increases in PI on the margins of the tropics (Fig. 1) appears to be consistent with this.

Even if future PI increases are accompanied by poleward shifts in LMI to the same extent as in the recent historical record, however, TC statistics will change, whether or not those changes are manifest in bulk statistics such as basin-wide ACE or PDI. Regions poleward of those where historical TC activity has been greatest would see substantial increases in activity. And at any fixed position, PI increases are still expected to lead to intensity increases for storms at that position, absent any other predictable countervailing change (e.g., increasing wind shear). The expectation of unprecedentedly strong future storms remains justifiable.

## Conclusions

We expect, based on a broad understanding informed by observations, theory, and numerical models, that tropical cyclone intensities should increase as the climate warms in response to human emissions of greenhouse gases. We have lower confidence, on the other hand, in recent historical trends in the actual and potential intensities of TCs (in basins other than the North Atlantic). These two conclusions are consistent with the idea that, in the absence of other systematic global influences, global tropical cyclone intensities (as measured by PI) are controlled by the large-scale environment.

PI trends themselves are subject to considerable uncertainty, with different observational data sets showing disagreement, particularly outside the North Atlantic. Climate model simulations indicate that during the late 20th century, greenhouse gas and aerosol radiative forcing had opposite and, until very recently, largely cancelling effects on PI. This cancellation renders radiatively forced trends in PI and actual intensity small and thus difficult to detect in the presence of large natural variability. Detection of trends in actual intensity is further complicated by data inhomogeneities and other observational limitations and by systematic poleward shifts in the storm tracks.

As the 21st century proceeds, we expect greenhouse gas warming to further outpace aerosol cooling and PI increases to exceed those observed to date. TC intensities at any given fixed location should increase accordingly, on average; simulations suggest trends on the order of 1 m $s^{-1}$ $decade^{-1}$ at the high end. If poleward shifts continue, these increases will be manifest in increases in activity at the poleward margins of TC basins, as well as in the occurrence of more intense storms (if perhaps fewer storms overall) in the historical cores of the basins.

These expectations could turn out to be incorrect because of unforeseen factors or inadequacies in our understanding. But the relatively low confidence in the detection of historical trends is not evidence of any such inadequacies.

## REFERENCES AND NOTES


1. S. B. Idso, R. C. Balling Jr., R. S. Cerveny, *Meteorol. Atmos. Phys.* **42**, 259–263 (1990).
2. A. J. Broccoli, S. Manabe, *Geophys. Res. Lett.* **17**, 1917–1920 (1990).
3. J. L. Evans, R. J. Allan, *Int. J. Climatol.* **12**, 611–623 (1992).
4. J. Lighthill *et al.*, *Bull. Am. Meteorol. Soc.* **75**, 2147–2157 (1994).
5. L. Bengtsson, M. Botzet, M. Esch, *Tellus* **48A**, 57–73 (1996).
6. T. R. Knutson *et al.*, *Nat. Geosci.* **3**, 157–163 (2010).
7. S. I. Seneviratne, N. Nicholls, D. Easterling, C. M. Goodess, S. Kanae, J. Kossin, Y. Luo, J. Marengo, K. McInnes, M. Rahimi, M. Reichstein, A. Sorteberg, C. Vera, X. Zhang, "Changes in climate extremes and their impacts on the natural physical environment," in *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change*, C. B. Field *et al.*, Eds. (Cambridge Univ. Press, 2012), pp. 109–230.
8. Intergovernmental Panel on Climate Change, *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, T. F. Stocker *et al.*, Eds. (Cambridge Univ. Press, 2013).
9. K. J. E. Walsh *et al.*, *WIREs Clim. Change* **7**, 65–89 (2016).
10. S. J. Camargo, A. A. Wing, *WIREs Clim. Change* **7**, 211–237 (2016).
11. M. Bister, K. A. Emanuel, *J. Geophys. Res.* **107**, 4801 (2002).
12. J. Persing, M. T. Montgomery, *J. Atmos. Sci.* **60**, 2349–2371 (2003).
13. R. K. Smith, M. T. Montgomery, S. Vogl, *Q. J. R. Meteorol. Soc.* **134**, 551–561 (2008).
14. G. H. Bryan, R. Rotunno, *Mon. Weather Rev.* **137**, 1770–1789 (2009).
15. G. A. Vecchi, B. J. Soden, *Nature* **450**, 1066–1070 (2007).
16. J. Yu, Y. Wang, K. Hamilton, *J. Clim.* **23**, 1354–1373 (2010).
17. S. J. Camargo, *J. Clim.* **26**, 9880–9902 (2013).
18. K. Emanuel, A. Sobel, *J. Adv. Model. Earth Syst.* **5**, 447–458 (2013).
19. L. M. Polvani, S. J. Camargo, R. R. Garcia, *J. Clim.* **29**, 2275–2289 (2016).
20. B. Tang, K. Emanuel, *J. Atmos. Sci.* **67**, 1817–1830 (2010).
21. I.-I. Lin *et al.*, *Geophys. Res. Lett.* **40**, 1878–1882 (2013).
22. P. Huang, I.-I. Lin, C. Chou, R.-H. Huang, *Nat. Commun.* **6**, 7188 (2015).
23. K. Emanuel, *J. Clim.* **28**, 8165–8170 (2015).
24. K. A. Emanuel, *Mon. Weather Rev.* **128**, 1139–1152 (2000).
25. A. A. Wing, A. H. Sobel, S. J. Camargo, *Geophys. Res. Lett.* **34**, L08810 (2007).
26. J. P. Kossin, S. J. Camargo, *Clim. Change* **97**, 329–337 (2009).
27. S. Wang, S. J. Camargo, A. H. Sobel, L. M. Polvani, *J. Atmos. Sci.* **71**, 4333–4348 (2014).
28. G. A. Vecchi, B. J. Soden, *Geophys. Res. Lett.* **34**, L08702 (2007).
29. M. A. Bender *et al.*, *Science* **327**, 454–458 (2010).
30. H. Murakami *et al.*, *J. Clim.* **25**, 3237–3260 (2012).
31. T. R. Knutson *et al.*, *J. Clim.* **28**, 7203–7224 (2015).
32. J. V. Manganello *et al.*, *J. Clim.* **27**, 7622–7646 (2014).
33. W. M. Frank, G. S. Young, *Mon. Weather Rev.* **135**, 3587–3598 (2007).
34. B. Wang, Y. Yang, Q.-H. Ding, H. Murakami, F. Huang, *Geophys. Res. Lett.* **37**, L07704 (2010).
35. D. J. Raymond, S. L. Sessions, Ž. Fuchs, *Q. J. R. Meteorol. Soc.* **133**, 1743–1754 (2007).
36. T. J. Dunkerton, M. T. Montgomery, Z. Wang, *Atmos. Chem. Phys.* **9**, 5587–5646 (2009).
37. K. A. Emanuel, D. S. Nolan, in *Proceedings of 26th Conference on Hurricanes and Tropical Meteorology* (American Meteorological Society, 2004), pp. 240–241.
38. K. Emanuel, *J. Adv. Model. Earth Syst.* **2**, 1 (2010).
39. M. K. Tippett, S. J. Camargo, A. H. Sobel, *J. Clim.* **24**, 2335–2357 (2011).
40. C. L. Bruyère, G. J. Holland, E. Towler, *J. Clim.* **25**, 8611–8626 (2012).
41. S. J. Camargo, A. H. Sobel, A. G. Barnston, K. A. Emanuel, *Tellus* **59A**, 428–443 (2007).
42. S. J. Camargo, M. K. Tippett, A. H. Sobel, G. A. Vecchi, M. Zhao, *J. Clim.* **27**, 9171–9196 (2014).
43. K. Emanuel, R. Sundararajan, J. Williams, *Bull. Am. Meteorol. Soc.* **89**, 347–367 (2008).
44. M. Wehner *et al.*, *J. Clim.* **28**, 3905–3925 (2015).
45. I. M. Held, B. J. Soden, *J. Clim.* **19**, 5686–5699 (2006).
46. K. A. Emanuel, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 12219–12224 (2013).
47. K. E. Trenberth, J. Fasullo, *J. Geophys. Res.* **112**, D23107 (2007).
48. G. Villarini *et al.*, *J. Clim.* **27**, 4622–4641 (2014).
49. E. Scoccimarro *et al.*, *J. Clim.* **27**, 4642–4654 (2014).
50. H.-S. Kim *et al.*, *J. Clim.* **27**, 8034–8054 (2014).
51. J. D. Woodruff, J. L. Irish, S. J. Camargo, *Nature* **504**, 44–52 (2013).
52. N. Lin, K. Emanuel, M. Oppenheimer, E. Vanmarcke, *Nat. Clim. Change* **2**, 462–467 (2012).
53. N. Lin, K. Emanuel, *Nat. Clim. Change* **6**, 106–111 (2015).
54. D. T. Resio, J. L. Irish, *Curr. Clim. Change Rep.* **1**, 74–84 (2015).
55. C. W. Landsea, B. A. Harper, K. Hoarau, J. A. Knaff, *Science* **313**, 452–454 (2006).
56. C. W. Landsea, G. A. Vecchi, L. Bengtsson, T. R. Knutson, *J. Clim.* **23**, 2508–2519 (2010).
57. G. A. Vecchi, T. R. Knutson, *J. Clim.* **24**, 1736–1746 (2011).
58. P. J. Klotzbach, C. W. Landsea, *J. Clim.* **28**, 7621–7629 (2015).
59. N. L. Bindoff, P. A. Stott, K. M. AchutaRao, M. R. Allen, N. Gillett, D. Gutzler, K. Hansingo, G. Hegerl, Y. Hu, S. Jain, I. I. Mokhov, J. Overland, J. Perlwitz, R. Sebbari, X. Zhang, "Detection and attribution of climate change: From global to regional," in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, T. F. Stocker *et al.*, Eds. (Cambridge Univ. Press, 2013), pp. 867–952.
60. P. J. Webster, G. J. Holland, J. A. Curry, H.-R. Chang, *Science* **309**, 1844–1846 (2005).
61. J. B. Elsner, J. P. Kossin, T. H. Jagger, *Nature* **455**, 92–95 (2008).
62. J. P. Kossin, T. L. Olander, K. R. Knapp, *J. Clim.* **26**, 9960–9976 (2013).
63. K. Emanuel, *Nature* **436**, 686–688 (2005).
64. K. Emanuel, *J. Clim.* **20**, 5497–5509 (2007).
65. G. A. Vecchi, K. L. Swanson, B. J. Soden, *Science* **322**, 687–689 (2008).
66. H. A. Ramsay, A. H. Sobel, *J. Clim.* **24**, 183–193 (2011).
67. S. J. Camargo, M. Ting, Y. Kushnir, *Clim. Dyn.* **40**, 1515–1529 (2013).
68. M. E. Mann, K. A. Emanuel, *EOS* **87**, 233, 238, 241 (2006).
69. B. B. B. Booth, N. J. Dunstone, P. R. Halloran, T. Andrews, N. Bellouin, *Nature* **484**, 228–232 (2012).
70. K. Emanuel, S. Solomon, D. Folini, S. Davis, C. Cagnazzo, *J. Clim.* **26**, 2288–2301 (2013).
71. G. A. Vecchi, S. Fueglistaler, I. M. Held, T. R. Knutson, M. Zhao, *J. Clim.* **26**, 3877–3891 (2013).
72. M. Ting, S. J. Camargo, C. Li, Y. Kushnir, *J. Clim.* **28**, 3926–3942 (2015).
73. G. Holland, C. Bruyère, *Clim. Dyn.* **42**, 617–627 (2014).
74. Materials and methods are available as supplementary materials on *Science* Online.
75. A. A. Wing, K. Emanuel, S. Solomon, *Geophys. Res. Lett.* **42**, 8669–8677 (2015).
76. M. Free, M. Bister, K. Emanuel, *J. Clim.* **17**, 1722–1727 (2004).
77. J. P. Kossin, *Bull. Am. Meteorol. Soc.* **96**, 1089–1096 (2015).
78. S. J. Smith *et al.*, *Atmos. Chem. Phys.* **11**, 1101–1116 (2011).
79. D. T. Shindell *et al.*, *Atmos. Chem. Phys.* **13**, 2653–2689 (2013).
80. B. Stevens, *J. Clim.* **28**, 4794–4819 (2015).
81. D. P. van Vuuren *et al.*, *Clim. Change* **109**, 5–31 (2011).
82. J. P. Kossin, K. A. Emanuel, G. A. Vecchi, *Nature* **509**, 349–352 (2014).
83. J. P. Kossin, K. A. Emanuel, S. J. Camargo, *J. Clim.* (2016).
84. S. M. Uppala *et al.*, *Q. J. R. Meteorol. Soc.* **131**, 2961–3012 (2005).


## ACKNOWLEDGMENTS

A.H.S. thanks J. Kossin, G. Vecchi, and K. Emanuel for helpful discussions. This work was supported by NSF grant AGS-1143959 and Office of Naval Research grant N00014-16-1-2073. A.A.W. is supported by a NSF Atmospheric and Geospace Sciences postdoctoral research fellowship under award no. 1433251.

## SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/353/6296/242/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S3
References (*85*–*87*)

10.1126/science.aaf6574

REVIEW

# Global trends in satellite-based emergency mapping

Stefan Voigt,[1]* Fabio Giulio-Tonolo,[2] Josh Lyons,[3] Jan Kučera,[4] Brenda Jones,[5] Tobias Schneiderhan,[1] Gabriel Platzeck,[6] Kazuya Kaku,[7] Manzul Kumar Hazarika,[8] Lorant Czaran,[9] Suju Li,[10] Wendi Pedersen,[11] Godstime Kadiri James,[12] Catherine Proy,[13] Denis Macharia Muthike,[14] Jerome Bequignon,[15] Debarati Guha-Sapir[16]

**Over the past 15 years, scientists and disaster responders have increasingly used satellite-based Earth observations for global rapid assessment of disaster situations. We review global trends in satellite rapid response and emergency mapping from 2000 to 2014, analyzing more than 1000 incidents in which satellite monitoring was used for assessing major disaster situations. We provide a synthesis of spatial patterns and temporal trends in global satellite emergency mapping efforts and show that satellite-based emergency mapping is most intensively deployed in Asia and Europe and follows well the geographic, physical, and temporal distributions of global natural disasters. We present an outlook on the future use of Earth observation technology for disaster response and mitigation by putting past and current developments into context and perspective.**

Disaster responders and the humanitarian community increasingly use Earth Observation (EO) satellite systems to assess the impact of and to plan and coordinate emergency response activities after major natural disasters around the world. EO systems provide response and relief workers with tools to lift the "fog off disaster." EO satellites help overcome operational uncertainties after major disasters that impede emergency response because of limited, incomplete, and often contradictory ground information. Furthermore, EO satellites provide emergency responders with a situational overview otherwise difficult to obtain during an ongoing disaster event. For example, synthetic aperture radar (SAR) sensors can see through storm clouds to remotely assess in near real time the exact extent or severity of flood disasters as they unfold. Local, national, and international agencies also use satellite-based emergency mapping (SEM) as part of larger resilience strategies (*1*) to help design, implement, and evaluate disaster risk reduction and recovery programs (*2*–*4*). The ultimate goal of SEM is to improve disaster relief effectiveness and thus to help reduce suffering and fatalities before, during, and after a disaster event occurs. We focus our Review on the response phase immediately after a disaster, which typically lasts from several days to a few weeks. This phase is technically challenging because of the strict time constraints and demands special skill sets and coordination among disaster responders, the SEM community, satellite operators, and international organizations. The global SEM response capabilities have been developing over the past 15 years and can today be considered to be at the forefront of the use of satellite technology and geoinformation in the broader field of disaster risk management (Box 1) (*5*).

> ***"The availability of ... EO satellite systems has increased during the past 15 years."***

Partly in response to growing demand, larger satellite constellations with more advanced sensors are being built, with the potential to provide unprecedented capacity for monitoring the Earth more rapidly and in more detail than ever before. This development has not been limited to the traditional space agencies in Europe, Japan, and the United States. Over the past 15 years, countries throughout Latin America, Africa, and Asia have started their own space programs. Dozens of new satellites have been launched, transforming availability and access to EO technology and data, further expanding the EO constellations and the ease of use of satellite data. The provision of vast quantities of raw satellite data to the disaster response community has no operational value per se. Being time sensitive in its relevance to immediate disaster mitigation, the data need to be rapidly processed, analyzed, and transformed by remote sensing professionals (*6*) into intuitive and understandable information products such as maps or reports; these can then directly be used in emergency management operations (*7*, *8*).

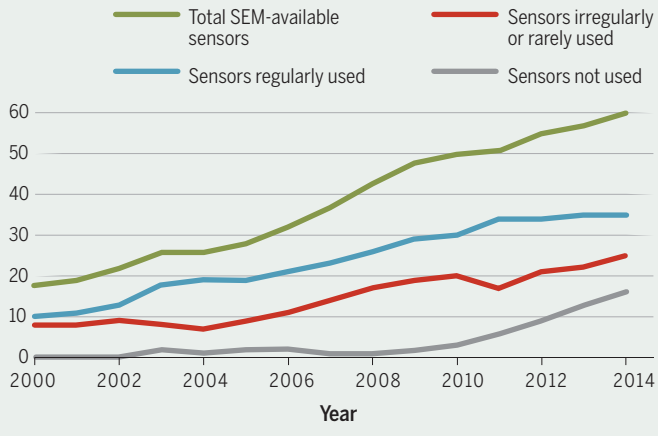
In reviewing global SEM responses of the past 15 years, five major events stand out, given their influence on the development of the international SEM community: (i) After the Indian Ocean Tsunami in 2004 (*7*), widespread international SEM cooperation and response coordination were necessary owing to the scale of the event, size of the impacted region, and the number of countries affected. During the disaster, satellite mapping played an important role by providing an overview of the situation on the ground and helping people to understand the magnitude of devastation caused by the tsunami. (ii) The Wenchuan Earthquake in 2008 (*9*) mobilized an at that time unprecedented number of programmed satellites and acquired satellite imagery for a single disaster event. Analysis and mapping of the data was mainly organized by the National Disaster Reduction Centre of China (NDRCC) and resulted in the generation of numerous satellite products. During this event, it became clear that satellite imagery alone could not suffice to assess more subtle structural earthquake damage to buildings and infrastructure. In response to this, the emergency-mapping community realized the need for airborne sensors and imagery from unmanned airborne vehicles (UAVs) in order to complement satellite-derived products. (iii) The Haiti Earthquake in 2010 (*10*) marked a turning point in the accessibility of openly licensed post-event satellite imagery to a broader internet and crisis-mapping community. Many satellite-based emergency maps were produced by many different organizations, which led to an overflow of SEM products and some criticism by the international disaster relief community (*11*). As a result, the International Working Group on Satellite-based Emergency Mapping (IWG-SEM) (*12*) was established to improve mutual information sharing, harmonization, and cooperation across the international SEM community. (iv) The Pakistan flood in 2010 (*13*, *14*) affected ~5% of the Indus River basin and 20 million people. Many varying SEM products were produced by different initiatives. The emergency response community was again overwhelmed with information, making it challenging to prioritize and ingest all the information into their operational workflows. The main concern was the thematic accuracy of the post-event information because of map products showing different extents of affected areas, such as the extent of flooding. This was another catalyst that led to the creation of the IWG-SEM (19, 20). (v) After The Great East Japan Earthquake in 2011 (Tohoku-Oki) (*15*), the Japan Aerospace Exploration Agency (JAXA) enlisted the help of international SEM

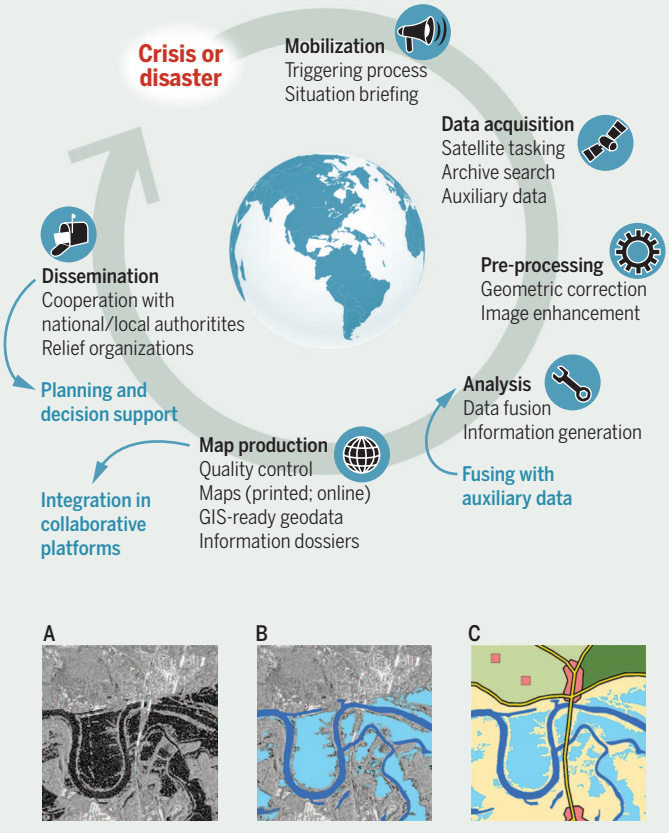
[1]German Aerospace Center, Oberpfaffenhofen, Germany. [2]Information Technology for Humanitarian Assistance, Cooperation and Action, Torino, Italy. [3]Human Rights Watch, Geneva, Switzerland. [4]European Commission - Joint Research Centre, Ispra, Italy. [5]U.S. Geological Survey, Sioux Falls, SD, USA. [6]Gulich Institute - Córdoba National University/CONAE, Córdoba, Argentina. [7]Japan Aerospace Exploration Agency, Tsukuba, Japan. [8]Asian Institute of Technology, Klong Luang, Pathumthani, Thailand. [9]United Nations Office for Outer Space Affairs, Vienna, Austria. [10]National Disaster Reduction Center of China Beijing, China. [11]Geneva International Centre for Humanitarian Demining, Geneva, Switzerland. [12]National Space Research and Development Agency, Abuja, Nigeria. [13]Centre National d'Études Spatiales, Toulouse, France. [14]Regional Centre for Mapping of Resources for Development, Nairobi, Kenya. [15]European Space Agency, Brussels, Belgium. [16]Université catholique de Louvain (UCL), Brussels, Belgium.
*Corresponding author. Email: stefan.voigt@dlr.de

**Box 1. Earth Observation satellites and principles in emergency mapping.**

The availability of scientific and commercial polar orbiting EO satellite systems has increased during the past 15 years (*41*–*44*). These satellites are equipped with imaging sensors in the visible and near- to mid-infrared part of the electromagnetic spectrum or in the radar frequencies. Systems useful for disaster extent and impact mapping have a ground sampling distance (GSD) in the range of 0.3 m to more than 300 m. A team of experienced image analysts can take from 6 to 16 hours to extract the relevant information from new satellite imagery and turn it into geo-information products, such as maps, for situation centers or decision-makers. Re-programming the satellite systems and collecting imagery over the disaster site typically takes 1 or 2 days and is considered one of the more time-consuming parts of the overall process (*6*). Many elements of the SEM production chain are becoming automated.

**Number of optical and radar satellite sensors (<300 MGSD)**




(**Above**) **The main steps involved in SEM product generation.** (**A**) Satellite image acquisition and preprocessing. (**B**) Image analysis and information extraction. (**C**) Visualization in dedicated geo-information products.

organizations immediately, contributing to efforts by national and local government bodies to collect information and support relief activities. By teaming up with different partners, more than 5000 satellite images were collected for assessment after the earthquake. These images were used to determine the overall extent of the damage and assess local conditions such as the availability of key facilities, allowing for the prioritization of the disaster response activities (*15*).

In the light of these experiences and the developments described above, the following questions may be raised: What are the temporal trends in overall SEM response rate and time? Are the SEM resources being deployed in the areas of greatest need? What is the individual reach of the different SEM mechanisms? Can we ascertain the fitness for purpose of the SEM? In order to address these questions, we systematically reviewed and assessed more than 1000 satellite emergency-mapping activations for natural and partially manmade disasters occurring globally between 2000 and 2014. The team captured SEM activation parameters of five relevant SEM mechanisms: the International CHARTER Space and Major Disasters (*16*–*18*), the European COPERNICUS program (including the phase when it was still called GMES) (*19*), United Nations (UNOSAT-UNITAR and ReliefWeb), SENTINEL ASIA (*20*), and the NDRCC (table S1). We linked the data with statistical ground-based information on the associated disaster events from the International Disaster Database EM-DAT (*21*) and data from the World Bank (*22*, *23*).

## Temporal trends in satellite response

To approach the first question, we investigated the temporal trends in SEM response. Globally, we observed a steady increase of SEM activities, growing from seven activations in 2000 to 123 activations in 2014 (Fig. 1A). This trend in international efforts in SEM is an encouraging sign of the readiness of this technology to support disaster management. It is likely that technological innovations—such as internal mechanism enhancements and the launch of virtual globes on the internet—has raised awareness and acceptance of geospatial data, leading to an increase in the number of disasters covered by SEM in the 2006 timeframe (Fig. 1B, dashed line). The COPERNICUS program seems to be the only SEM mechanism still increasing the number of disaster events analyzed per year, whereas other SEM mechanisms seem to have maximized in activation numbers from around 2010 on. The COPERNICUS program is strongly supported by the European Commission policy and funding, with operational integration into the European Union member state administration and disaster management procedures just beginning. Thus, this program is also expected to grow in the years to come. Furthermore, the average number of mapping products has increased from two to five products per SEM activation between 2000 and 2014 (Fig. 1C). This suggests that the SEM community has substantially expanded the capacity to turn satellite imagery into geo-information and mapping products for disaster response purposes. This also implies a greater need for well-organized cooperation, harmonization, and product standardization at a global scale in order to make more coherent use of the international space and ground-based capacities.

Similarly, we assessed delay times from mobilization to image availability so as to understand SEM responsiveness. We controlled for shared activations between SEM mechanisms and excluded nonrapid SEM responses [time ($T$) > 1 week]. The time series analysis on responsiveness only started in 2001 because consistent and reliable records became available at that time. In 2006, the average overall response time from mobilization to first product was ~4.5 days;

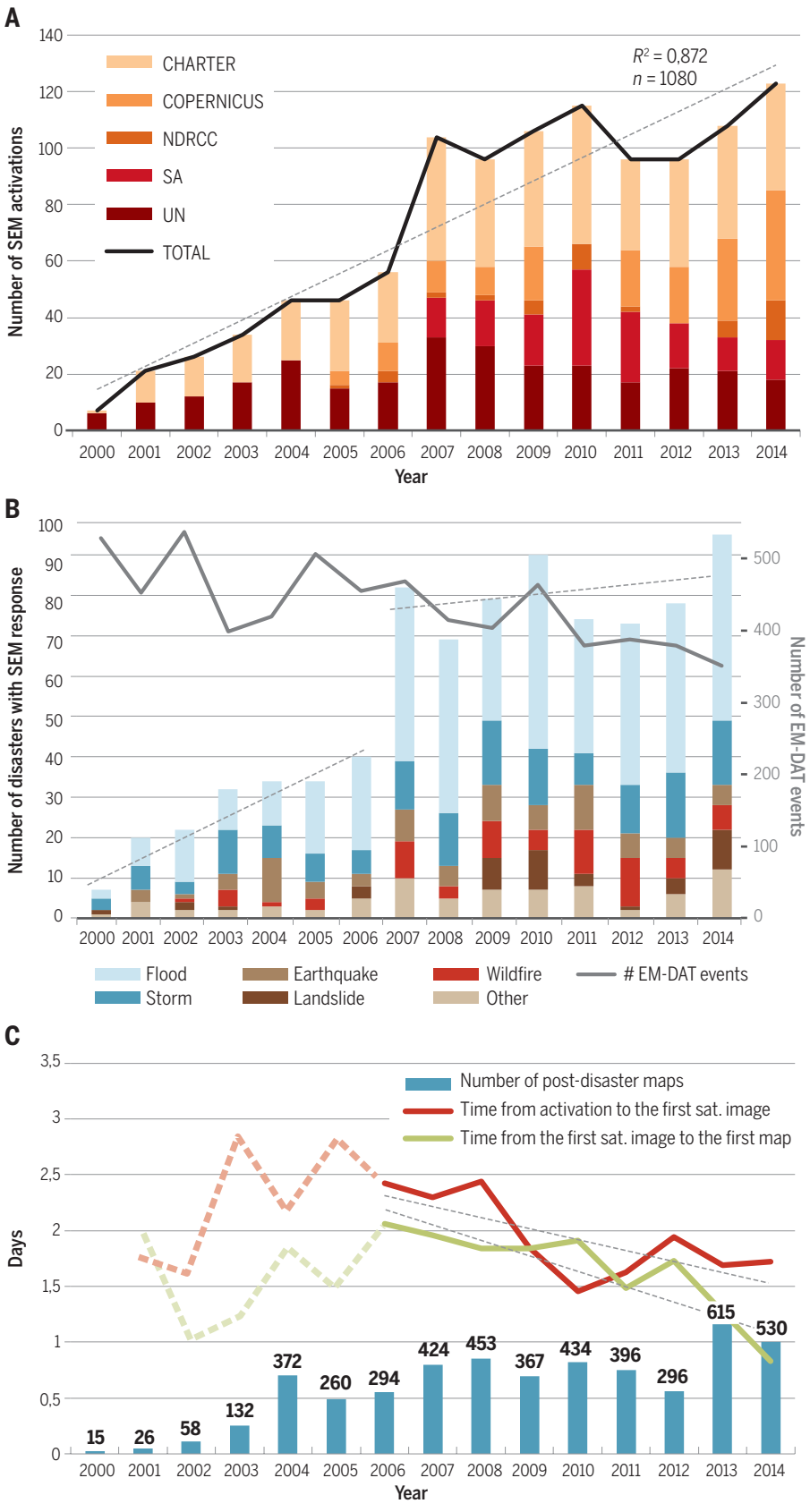



**Fig. 1. SEM temporal trends during the study period 2000–2014.** (**A**) Number of activations and distribution among the different SEM mechanism. (**B**) The differentiation by disaster types over time. (**C**) The overall map production volumes and response times.

this was reduced to ~2.5 days on average by 2014 (Fig. 1C).

## Global spatial pattern in SEM

To determine whether the SEM resources were deployed in areas of greatest need, we visualized the global spatial patterns of SEM responses ($n$ = 804 geocoded SEM responses), controlling for shared activations and synoptically displaying a global population density data set (Fig. 2) (*24*, *25*). The analysis indicates that a large majority (~75%) (fig. S1) of SEM activations are related to hydrometeorological disasters (similar to the EM-DAT percentages for the same period), which cluster in distinct parts of the Americas, Africa, eastern/southeastern Asia, and Europe. However, the SEM activations related to geophysical disasters are clustered along the borders between the Nazca and South American plates as well as the Eurasian and Indo-Australian plates. The main mixed clusters of hydrometeorological and geophysical SEM activities are located in the European and Himalayan regions as well as in parts of the Northern Andes, Central America, and the Caribbean. Generally, the locations of SEM activations resemble the large global natural hazards patterns, with the seismic active zones and the major storm systems. In Europe, the distribution of SEM activations for geophysical events is situated along the border between the African and the Eurasian plates (Fig. 2B). We identified a spatial correlation between the location of the SEM activities and the densely populated areas of the world. This suggests that human exposure drives the decision to request international satellite emergency mapping. We observed that for main parts of India, international SEM activities are not being called on, which is likely to be explained by a preference for domestic SEM capacities in this region (Fig. 2C).

The north-south distribution of the SEM activations matches well with the relative distribution of global population (Fig. 3A). Deviations occur at 15°N and 20°S, with frequent cyclonic flood and storm events hitting at these latitudes. Another deviation occurs around 25°N; we attribute this to the high population density and relatively few international SEM activations over the Indian subcontinent. As for the relative distribution of the land masses, the low populated Nordic landmasses (Canada/Russia/Alaska) and southern sparsely populated regions in the Amazonas, southern Africa, and Australia (mainly between 5°N and 35° S) are not subject to substantial numbers of SEM activities.

Generally, of all studied SEM activations, 25% cover six countries (EM-DAT 25%, six countries), 50% cover 21 countries (EM-DAT 50%, 25 countries), and 75% cover 50 countries (EM-DAT 75%, 62 countries), whereas the overall SEM activations have reached 163 countries.

We observed variations in the use of the SEM mechanisms as well as the distribution of disasters covered by means of SEM. Among the 12 countries with the largest disaster occurrences (according to EM-DAT), Pakistan and Vietnam managed 26 and 19% of their domestic disasters, respectively,

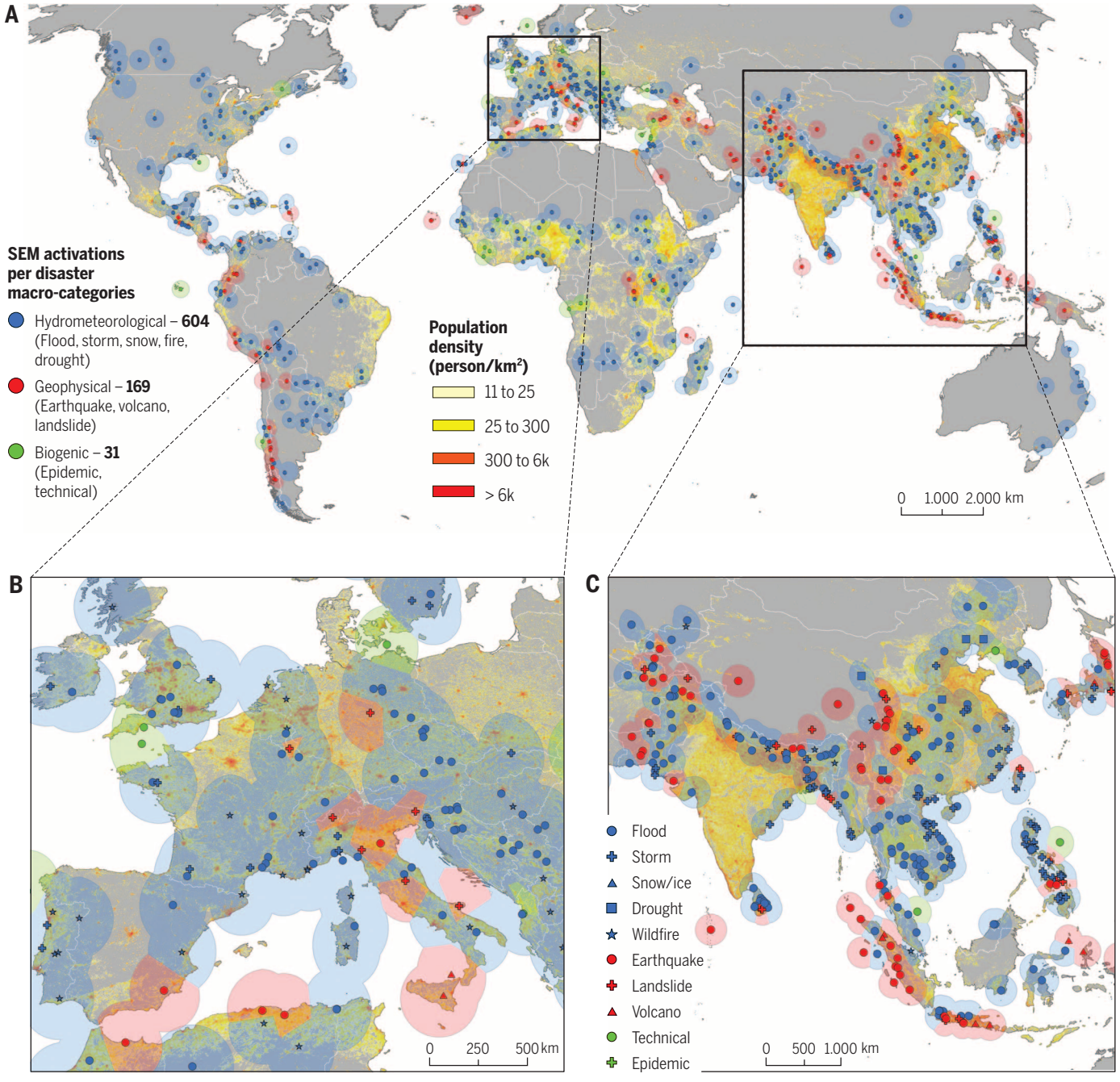



**Fig. 2. Spatial distribution of SEM activations by disaster type.** (**A**) At the global level. The distribution of SEM activations are grouped according to three disaster macrocategories: (i) hydrometeorological, blue symbols (including flood, storm, snow, wildfire, and drought events); (ii) geophysical, red symbols (earthquake, volcano, and landslide events); and (iii) biogenic, green symbols (epidemic outbreaks and technical accidents). (**B** and **C**) At regional level. (B) Western and southern Europe. (C) Southern, eastern, and southeastern Asia. The detailed disaster types can be read from the individual symbols. Polygons highlight the clustering of activations aggregated at disaster macrocategory level. All three sections show population density in the background (*24*).

with SEM support. Countries such as China, India, Philippines, Indonesia, Bangladesh, and Japan managed between 10 and 15% of their domestic disasters with SEM. The United States, Afghanistan, Mexico, and Russia range between 5 and 8%. We also found that Asia is the main global focus of international SEM activities, which is in line with the fact that according to EM-DAT, more disasters occur in this region as compared with others. The CHARTER was activated by the United States more than by any other country, and COPERNICUS was activated mainly for disaster situations in southern and southeastern Europe (table S2). In almost all regions of the world, SEM activities have risen in number substantially during the past 5 years. Only for the Americas and the Caribbean has the SEM frequency remained stable or slightly decreased during the past 5 years. Eastern and Western Africa have also remained stable, with a relatively high level of SEM activities over the past 10 years, whereas Australia, Polynesia, and Melanesia are covered by only a few SEM activations during the study period (Fig. 3B).

### The reach of individual SEM mechanisms

The CHARTER, because of its global scope and as supported by its recent universal access efforts, is the most widely active and fully international SEM

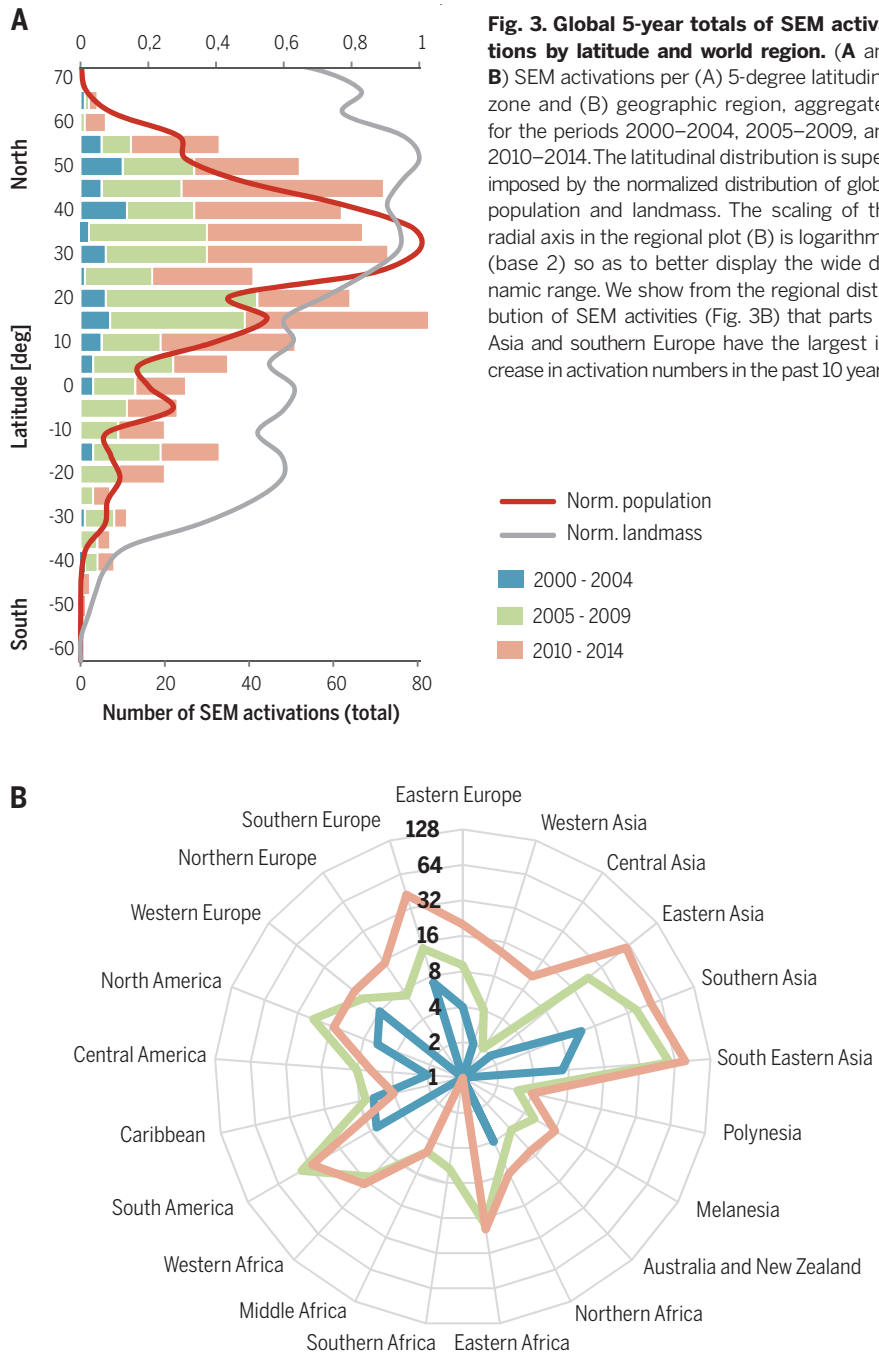



**Fig. 3. Global 5-year totals of SEM activations by latitude and world region.** (**A** and **B**) SEM activations per (A) 5-degree latitudinal zone and (B) geographic region, aggregated for the periods 2000–2004, 2005–2009, and 2010–2014. The latitudinal distribution is superimposed by the normalized distribution of global population and landmass. The scaling of the radial axis in the regional plot (B) is logarithmic (base 2) so as to better display the wide dynamic range. We show from the regional distribution of SEM activities (Fig. 3B) that parts of Asia and southern Europe have the largest increase in activation numbers in the past 10 years.

mechanism. In North America, the CHARTER is the only SEM mechanism used to complement domestic capacities. In South America, Africa, Europe, and Asia, the SEM activities of the CHARTER are increasingly complemented by other mechanisms. For Europe, the COPERNICUS Emergency Management Service (EMS) plays an increasingly important role, resulting in a substantial decrease of CHARTER activations over Europe in the past 5 years. In Asia, the SENTINEL ASIA activities have a strong prevalence, accompanied by CHARTER and United Nations (UN) mapping efforts. The latter also plays a major role in Africa, the Near East, Central America, the Caribbean, and South America. For China, the NDRCC plays the most important role, with the CHARTER complementing NDRCC capacities. The top five ranking countries for which each SEM mechanism was activated is provided in table S2.

## Is global SEM fit for the purpose?

This question is challenging and will require further research to provide a quantitative answer. There have been attempts to assess the value of space- and geo-information for disaster and risk management, including economic and operational value (*26*, *27*). At this time, these assessments remain qualitative. To find a quantitative answer or trends in the usefulness of SEM over time, many complex technological trends (such as impact of information and communication technology, awareness, and preparedness-raising) would have to be eliminated or controlled for. Generally, the demand for and quality of services of global SEM activities has risen over the past decade. In many countries, strong efforts are being made to build SEM capacities within the emergency management services (*8*, *15*), suggesting that they are useful, although the benefits cannot yet be quantified in absolute numbers.

As SEM products increase in quantity, timeliness, and complexity, there are other kinds of emergencies and disaster phases that may benefit. It has already been demonstrated that satellite analyses can be highly relevant to slow on-set events with a vast geographical impact, such as water scarcity or drought (*28*–*30*). Even geodetic satellite signals combined with solid Earth load modeling can be used for drought pattern and severity estimation (*31*). Also, the monitoring of associated large mass population movements, which are difficult to track on the ground, could benefit from existing SEM capabilities (*32*). There are global communities that monitor crop yield and drought by means of satellites (*33*, *34*); however, they are not yet well integrated into the SEM community. The 2011 Horn of Africa drought is an example of a missed opportunity for better integration and consolidation of SEM into global mechanisms aimed at boosting targeted response (*35*). In 2015, SEM was enlisted in response to the Ebola crisis, demonstrating its potential for supporting global health crises (*36*). Of course, SEM can only be used to monitor health-related parameters and indirect physical consequences of epidemic or pandemic situations on the Earth's surface. However, during the Ebola crisis, SEM was used intensively for mapping and planning of health posts as well as mapping of oil palm trees because fruit bats are considered the main natural host of the virus (*37*). After the magnitude 7.8 Gorkha earthquake in Nepal in 2015, extensive use of ad hoc satellite image collections was made, going beyond the regular rapid SEM response, for surveying geohazards such as landslides and destabilized glacier lakes in the mountainous and remote regions that are difficult to assess otherwise (*38*).

## Conclusion

The comparison between EM-DAT and SEM distributions indicates that global SEM activities are progressively evolving. However, rapid response, accuracy, and increased frequency of SEM mappings are necessary considering the growing vulnerability of global societies, technological dependencies, and projected climate change scenarios. Therefore, the scope of global SEM activities should be broadened to better include drought, extreme temperature events, global pandemics, and other slow on-set events. Nonetheless, a major challenge for EO disaster response is still the satellite tasking, reprogramming,

and image collection; these require ~2 days on average to complete, as compared with the ~6 to 8 hours required for mapping after the availability of satellite imagery.

Generally speaking, 30 years after the UN General Assembly resolution on remote-sensing principles and the pledge that "Remote sensing shall promote the protection of mankind from natural disasters" (*39*), the initial organizational and procedural hurdles for making satellite analysis available for operational disaster management have been mostly overcome. Recognizing the diversification and intense utilization of SEM at a global scale, we suggest the establishment of international guidelines on emergency mapping, quality assurance, and harmonization, tailored to specific disaster types. In addition, operational global partnerships among agencies and organizations are essential for strengthening space-based disaster relief efforts. Cooperation among the operational SEM mechanisms must be intensified, within the IWG-SEM (*40*), and UN-SPIDER (*1*), as well as through other regional and global initiatives. Improved real-time information exchange on SEM activities, mapping requirements, and locations of available SEM-derived products at any given time is a key step in this process.

In the coming years, government, public, and commercial sectors will have greater capacity for imaging through satellite constellations, such as the European Copernicus Sentinel constellation and the many commercial systems with very-high-resolution optical imaging capability that are in operation or coming up. With the higher throughput of large quantities of imagery and increasingly higher spatial resolution of satellite data, automation and image data mining as well as mass-data processing techniques will play a key role in the global SEM landscape. Single images for disaster mapping will hand over to multiscale, multitemporal nested monitoring approaches, which are relevant to identify disaster hotspots. Coarser and more frequent satellite imagery will be used to identify areas of concern and to then dynamically "zoom in" on the critical regions by using high-spatial-resolution image data. Near real-time observations and direct monitoring of dynamic natural disaster processes such as lava flows, landslides, or floods will be possible from space. In the next 5 to 10 years, substantial scientific, technological, and operational development will handle mass data from different satellite constellations and innovative space sensors. In addition, data relay satellites will be used for boosting reprogramming as well as data downlink. Moreover, automated pattern and object recognition from oblique observations of disaster scenarios is likely to come into wider use. The use of video sequences from space for disaster situation assessment and real-time processing and analysis of satellite imagery for visual analytics and fusion with crowd-sourced and social media information is also likely to play a bigger role, along with high-resolution geostationary EO systems for disaster situational awareness. Online imagery access services and geospatial big data platforms will further shape and advance the global SEM efforts in the near future. These technologies will not all develop at the same pace; nevertheless, there are substantial procedural changes and technological innovations in progress that should be used diligently in order to further advance the global SEM capacities in the years to come.

## REFERENCES AND NOTES


1. United Nations, Space-based Information for Disaster Management and Emergency Response, UN-SPIDER Knowledge Portal, accessed 10 December 2015; www.un-spider.org.
2. O. M. Bello, Y. A. Aina, *Procedia Soc. Behav. Sci.* **120**, 365–373 (2014).
3. United Nations Development Programme (UNDP), "Human Development Report" (UNDP, New York, USA, 2015).
4. Committee on Earth Observation Satellites, *Satellite Earth Observations in Support of Disaster Risk Reduction*, S. W. Ivan Petiteville *et al.*, Eds. (CEOS, 2015), pp. 84.
5. A. Aitsi-Selmi *et al.*, *Int. J. Disast. Risk Sci.* **7**, 1–29 (2016).
6. A. Ajmar, P. Boccardo, F. Disabato, F. Giulio Tonolo, *Rendiconti Lincei* **26** (suppl. 1), 63 (2015).
7. S. Voigt *et al.*, *IEEE Trans. Geosci. Remote* **45**, 1520–1528 (2007).
8. P. Boccardo, F. Giulio Tonolo, in *Engineering Geology for Society and Territory* (Springer, 2015), vol. 5, chap. 17, pp. 7.
9. X. Tong *et al.*, *ISPRS J. Photogramm. Remote Sens.* **68**, 13–27 (2012).
10. K. Duda, B. K. Jones, *Photogramm. Eng. Remote Sensing* **77**, 899–907 (2011).
11. S. Voigt *et al.*, *Photogramm. Eng. Remote Sensing* **77**, 923–931 (2011).
12. International Working Group on Satellite-Based Emergency Mapping (IWG-SEM) Webpage, accesseded 10 February 2016; http://iwg-sem.org.
13. K. Gaurav, R. Sinha, P. K. Panda, *Nat. Hazards* **59**, 1815–1826 (2011).
14. W. Bank, Pakistan Floods 2010—Preliminary damage and needs assessment, accessed 31 May 2016 (2010); www.gfdrr.org/sites/gfdrr/files/publication/Pakistan_DNA_0.pdf.
15. K. Kaku, N. Aso, F. Takiguchi, *Int. J. Disast. Risk Reduct.* **12**, 134–153 (2015).
16. J. L. Bessis, J. Bequignon, A. Mahmood, *Acta Astronaut.* **54**, 183–190 (2004).
17. A. Mahmood, J. L. Bessis, J. Bequignon, L. Lauritson, K. V. Venkatachary, *Int. Geosci. Remote Sens.* **2**, 771–773 (2002).
18. B. K. Jones, T. S. Striker, A. Mahmood, G. R. Platzeck, in *Time-Sensitive Remote Sensing*, C. D. Lippitt *et al.*, Eds. (Springer Verlag, 2015), pp. 79–89.
19. Regulation (EU) No 377/2014 of the European Parliament and of the Council of the 3 April 2014 establishing the Copernicus Programme and repealing Regulation (EU) No 911/2010 (2014); http://copernicus.eu/sites/default/files/library/Regulation_377_2014_Copernicus_3April2014.pdf.
20. K. Kaku, A. Held, *Int. J. Disast. Risk Reduct.* **6**, 1–17 (2013).
21. D. Guha-Sapir, P. Hoyois, R. Below, "Annual Disaster Statistical Review 2014—The Numbers and Trends," [Center for Research on the Epidemiology of Disasters (CRED), 2015].
22. World Bank, Country Income Groups, accessed January 15, 2016; http://data.worldbank.org/news/new-country-classifications-2015.
23. Materials and methods are available as supplementary materials on *Science* Online.
24. Oak Ridge National Laboratory, LandScan Population Data 2010, accessed 5 December 2015; http://web.ornl.gov/sci/landscan.
25. J. E. Dobson, E. A. Bright, P. R. Coleman, B. L. Bhaduri, in *Remotely Sensed Cities*, V. Mesev, Ed. (Taylor & Francis, London, Great Britain, 2003), pp. 267–281.
26. I. Ayala *et al.*, "The Value of Geoinformation for Disaster and Risk Management (VALID)," (Joint Board of Geospatial Information Societies, 2013).
27. Organisation for Economic Co-operation and Development, *OECD Handbook on Measuring the Space Economy* (OECD Publishing, 2012).
28. M. Rodell, I. Velicogna, J. S. Famiglietti, *Nature* **460**, 999–1002 (2009).
29. J. S. Famiglietti *et al.*, *Science* **349**, 2 (2015).
30. A. AghaKouchak *et al.*, *Rev. Geophys.* **53**, 452–480 (2015).
31. A. A. Borsa, D. C. Agnew, D. R. Cayan, *Science* **345**, 1587–1590 (2014).
32. F. Checchi, B. T. Stewart, J. J. Palmer, C. Grundy, *Int. J. Health Geogr.* **12**, 4–12 (2013).
33. F. Rembold, C. Atzberger, I. Savin, O. Rojas, *Remote Sens.* **5**, 1704 (2013).
34. C. Yang, J. H. Everitt, J. M. Bradford, *Precis. Agric.* **10**, 292–303 (2009).
35. R. Bailey, Managing famine risk—Linking early warning to early action, accessed 5 June 2016 (2013); www.chathamhouse.org/sites/files/chathamhouse/public/Research/Energy,%20Environment%20and%20Development/0413r_earlywarnings.pdf.
36. United Nations Office for the Coordination of Humanitarian Affairs, ReliefWeb—West Africa: Ebola Outbreak—Mar 2014, accessed 5 June 2016 (2014); http://reliefweb.int/disaster/ep-2014-000041-gin.
37. I. C. S. M. Disasters, Ebola Epidemic in West Africa, accessed 15 March 2016; www.disasterscharter.org/web/guest/activations/-/article/other-in-sierra-leone.
38. J. S. Kargel *et al.*, *Science* **351**, aac8353 (2016).
39. United Nations, 41/65 Resolution Relating to Remote Sensing of the Earth from Space, accessed 10 February 2016; www.un.org/documents/ga/res/41/a41r065.htm.
40. International Working Group on Satellite-Based Emergency Mapping, *Emergency Mapping Guidelines* (2014); www.un-spider.org/sites/default/files/IWG_SEM_EmergencyMappingGuidelines_v1_Final.pdf.
41. A. S. Belward, J. O. Skøien, *ISPRS J. Photogramm. Remote Sens.* **103**, 115–128 (2015).
42. University of Twente, ITC's Database of Satellites and Sensors, accessed 15 January 2016; www.itc.nl/research/products/sensordb/AllSatellites.aspx.
43. European Space Agency, Earth Observation Portal, accessed 1 December 2015; https://directory.eoportal.org/web/eoportal/satellite-missions.
44. Committee on Earth Observation Satellites, The CEOS Database, accessed 20 December 2015; http://database.eohandbook.com/timeline/timeline.aspx.


## ACKNOWLEDGMENTS

Data sources for this work include International Charter Space and Major Disasters (disasterscharter.org), the European Copernicus EMS (emergency.copernicus.eu), Sentinel Asia (sentinel.tksc.jaxa.jp), UNITAR/UNOSAT (unitar.org/unosat), UNOCHA/ReliefWeb (reliefweb.int), National Disaster Reduction Center of China (jianzai.gov.cn), CRED-EMDAT (emdat.be) Deutsches Zentrum für Luft- und Raumfahrt (DLR) Center for Satellite Based Crisis Information (zki.dlr.de), Sertit (sertit.u-strasbg.fr), and the World Bank (data.worldbank.org/news/new-country-classifications-2015). The support of the International Charter, Copernicus, Sentinel Asia, and NDRCC in granting access to additional SEM statistics and metadata for this Review is kindly acknowledged. We thank A. Thiele, V. Hertel, M. Juessi (DLR), T. Nakao (Asian Disaster Reduction Center), and G. Chaohui (United States Geological Survey) for their enthusiastic efforts in researching the data sources and in database compilation. Many thanks also to R. Kiefl (DLR), A. Vanderveken and A. Beirekdar (UCL), without whom database and geospatial processing would have been difficult. INVAP's support to facilitate the contribution of Gulich Institute to the Review is kindly acknowledged. Last, we thank E. Schoepfer and A. Mager (DLR) for support with graphics and G. Strunz (DLR) for carefully reviewing the manuscript.

## SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/353/6296/247/suppl/DC1
Materials and Methods
Fig. S1
Tables S1 and S2

10.1126/science.aad8728

REVIEW

# Connecting slow earthquakes to huge earthquakes

Kazushige Obara* and Aitaro Kato

Slow earthquakes are characterized by a wide spectrum of fault slip behaviors and seismic radiation patterns that differ from those of traditional earthquakes. However, slow earthquakes and huge megathrust earthquakes can have common slip mechanisms and are located in neighboring regions of the seismogenic zone. The frequent occurrence of slow earthquakes may help to reveal the physics underlying megathrust events as useful analogs. Slow earthquakes may function as stress meters because of their high sensitivity to stress changes in the seismogenic zone. Episodic stress transfer to megathrust source faults leads to an increased probability of triggering huge earthquakes if the adjacent locked region is critically loaded. Careful and precise monitoring of slow earthquakes may provide new information on the likelihood of impending huge earthquakes.

Megathrust earthquakes occurring along the boundary between a subducting oceanic plate and an overriding continental plate may cause devastating damage due to the resulting strong ground motion and possible tsunamis. If we can retrieve a precursory signal prior to a megathrust earthquake, it would be useful for disaster mitigation. Although many studies have investigated earthquake prediction, no practical method has been established for the short-term prediction of an impending huge earthquake. Even though long-term changes in $b$ value [which characterizes the magnitude-frequency distribution of seismicity (*1*)] and in the tidal sensitivity of regular earthquakes (*2*) were retrospectively observed in the mainshock rupture areas of the 2004 $M_w$ 9.2 Sumatra and 2011 $M_w$ 9.0 Tohoku earthquakes, we cannot yet come to any conclusion about the quantitative predictive power of these parameters (*3*, *4*).

Slow earthquakes show an intermediate type of fault slip that is transitional between the fast rupture of regular earthquakes and stable sliding along a megathrust fault interface. The deployment of geodetic and seismic observation networks such as GEONET (*5*) and Hi-net (*6*) in Japan has contributed to the discovery of various types of slow earthquakes. Slow earthquakes provide phenomenological evidence for the existence of a transition zone between locked and creeping zones proposed by thermal modeling studies (*7*) and the partial release of slip deficit. In the past two decades, slow earthquakes have been detected in many subduction zones along the Pacific Rim (*8*).

The existence of different modes of fault slip is useful for understanding the physics of earthquake phenomena. Furthermore, the characteristic activity of slow earthquakes might be linked to huge earthquakes, despite the limited period during which observations of such phenomena have been made. Soon after the discovery of slow earthquakes, it was suggested that slow earthquakes could lead to an increased probability of megathrust events (*9*), although there have been few observations to confirm or refute this hypothesis statistically. On the basis of observations accumulated during the past decade, we focus here on how the study of slow earthquakes can help to further our understanding of the earthquake cycle, which should in turn help to narrow and focus our ability to improve disaster preparedness for megathrust earthquake hazards.

Earthquake Research Institute, University of Tokyo, Bunkyo, Tokyo 113-0032, Japan.
*Corresponding author. Email: obara@eri.u-tokyo.ac.jp

## Types of slow earthquakes

Slow earthquakes in the Nankai subduction zone are typically categorized as seismic or geodetic events according to their characteristic time scale (Fig. 1). Geodetic slow earthquakes are characterized as being either long-term slow slip events (SSEs) with durations of months or years, or short-term SSEs with durations of days. Seismic slow earthquakes are characterized by very-low-frequency (VLF) earthquakes at dominant periods of tens of seconds, and by low-frequency tremor at dominant frequencies of several Hz. These slow earthquakes are located both shallower and deeper than the seismogenic zone of the 1946 $M_w$ 8.3 Nankai and 1944 $M_w$ 8.1 Tonankai earthquakes (Fig. 2). Deep tremor is nearly continuosly distributed within a narrow belt–like zone along the downdip edge of the megathrust seismogenic zone over a length of 600 km; the belt width is several tens of kilometers (*10*). Tremor episodes usually continue for several days and accompany both short-term SSEs (*11*) and deep VLF earthquakes (*12*). The association of short-term SSEs with tremor was first discovered in Cascadia (*13*, *14*); these coupled phenomena are also referred to as episodic tremor and slip (ETS).

In southwest Japan, long-term SSEs shift to shallower depths than the deep ETS zone, appearing to fill a gap between tremor and shallower locked zones (*15*). In the Tokai region near the eastern edge of the deep ETS zone, a long-term SSE occurred from 2000 to 2005, releasing a seismic moment equivalent to a $M_w$ 7.1 earthquake (*16*). In the Bungo channel at the westernmost part of the deep ETS zone, a $M_w$ 6.8 long-term SSE with a duration of several months occurs about every 6 or 7 years, accompanying tremor activity (*17*). On the shallower side of the locked zone, shallow VLF earthquakes are sporadically distributed on the landward side of the Nankai trough (*18*). Data from seafloor seismometers reveal that the intensification of shallow tremor was coincident with shallow VLF earthquake activity off the east coast of Kyushu (*19*). The shallow VLF earthquakes extend to the Ryukyu Islands from off Kyushu (*20*) and are associated with short-term SSEs (*21*). Another shallow VLF earthquake zone is located east of Hokkaido (*22*), strongly modulated by afterslip following the 2003 $M_w$ 8.1 Tokachi-oki earthquake.

In contrast to ETS observed in the Nankai subduction zone, a series of SSEs of $M_w$ ~6.6 located near the Boso Peninsula, central Japan, has consistently accompanied swarms of regular earthquakes on the downdip part of the SSE area (*23*). Each SSE, with a duration of weeks, occurs once every 5 to 7 years.

Among the various slow earthquake types observed along the Pacific Rim (Fig. 3), ETS behavior in Cascadia is similar to that seen in Nankai; however, the scale of activity is much larger in Cascadia, where the ETS zone extends for 1200 km along the strike of the subducting plate and is divided into several segments (*24*). Recently, deep VLF earthquakes were detected (*25*) but no accompanying long-term SSEs or shallow slow earthquakes were observed. In Mexico, long-term SSEs with $M_w > 7$ were detected first, followed by tremor in the downdip portion of the long-term SSE source area (*26*). Slow earthquakes around the North Island of New Zealand show more complicated behavior (*27*). That is, SSEs with durations of weeks occur at intervals of a few years along the shallow interplate zone of the Hikurangi margin accompanying regular earthquakes, similarly to the Boso SSEs. Recently, seafloor pressure gauge observations revealed that the slip area of the SSE extends close to the trench (*28*). At depths of 20 to 70 km, long-term SSEs occur with durations of several years, and minor tremor activity has been detected near the SSE-generating fault.

Tremor occurs not only in subduction zones but also along transform plate boundaries, such as at the deep extension of the San Andreas Fault (*29*). Tremor associated with teleseismic surface waves has also been detected in the zone of collision between the Pacific and Eurasian plates in southern Taiwan (*30*).

## Slow earthquakes as analogs of megathrust earthquakes

Megathrust earthquakes are characterized by their variability in size and recurrence interval (*31*). Nankai megathrust earthquakes seem to occur repeatedly; however, the rupture area and recurrence interval are highly variable. On the other hand,

some slow earthquakes are characterized by their regular, periodic occurrence. Typical recurrence intervals of slow earthquakes range from months to years, much shorter than those of megathrust earthquakes. Therefore, the repetitious nature of slow earthquakes may be useful for improving our understanding of rupture styles and recurrence cycles of megathrust earthquake.

Among the various types of slow earthquakes, ETS is analogous to megathrust earthquakes, as both exhibit recurrent activity and rarely rupture multiple fault segments. An ETS zone at the deeper extension of the seismogenic zone along a plate interface can be separated into segments (Fig. 4) according to recurrence interval, which ranges from 3 to 6 months in southwest Japan (*11*) and from 10 to 19 months in Cascadia (*24*). This recurrence interval is more regular than those for megathrust earthquakes, although the interval and rupture area for each episode are variable. ETS events usually migrate within each segment and rarely propagate to adjacent segments. For example, tremor episodes occur independently at intervals of 6 months in the Kii and Tokai segments, separated by a gap at Ise Bay (Fig. 2). However, during the 2006 ETS episode, SSEs continuously propagated from the Kii to the Tokai segments through the Ise Bay gap, without any excitation of tremor (*12*). The continuous propagation of minor slow slip without tremor has also been observed at Cascadia (*32*). The gap zone does not always host tremor, allowing smooth migration of tremor on other occasions. These observations suggest that frictional properties may change over space and time at the gaps that characterize segment boundaries. ETS migration that extends to an adjacent segment is analogous to multisegment rupture during huge earthquakes. Consequently, ETS zone segment boundaries play an important role in controlling rupture termination.

During the 2003 and 2010 episodes in the Bungo channel, long-term SSE recurred with almost identical slip parameters, including moment magnitude, slip area, and duration of SSE (*33*). This SSE seems to be a characteristic earthquake, which is a unique earthquake rupturing at approximately the same source with a regular interval. By contrast, the moment magnitude and slip area of each Boso SSE were not always identical, as determined from the longest observed repetition history of any SSE, almost 30 years. The four most recent SSEs detected by GNSS (Global Navigation Satellite System) occurred in 2002, 2007, 2011, and 2014, indicating a gradual shortening of the recurrence interval, possibly due to stress transfer related to the 2011 Tohoku earthquake (*23*, *34*). Alternatively, this shortening may signal a stress-state change caused by an approaching large earthquake, as seen in numerical simulations (*35*) and discussed below. On the other hand, a small SSE was detected by seismological means immediately after the 2011 Tohoku earthquake (*36*). If we take this small SSE hidden by substantial afterslip following the Tohoku earthquake into account, the recurrence interval shows a more complex pattern, and the entire recurrence history must be reconsidered.

## Slow earthquakes as stress meters

Slow earthquakes are sensitive to external stress perturbations because of fault weakness. During ETS, tremor is usually modulated by stress variations associated with Earth tides (*37*); during

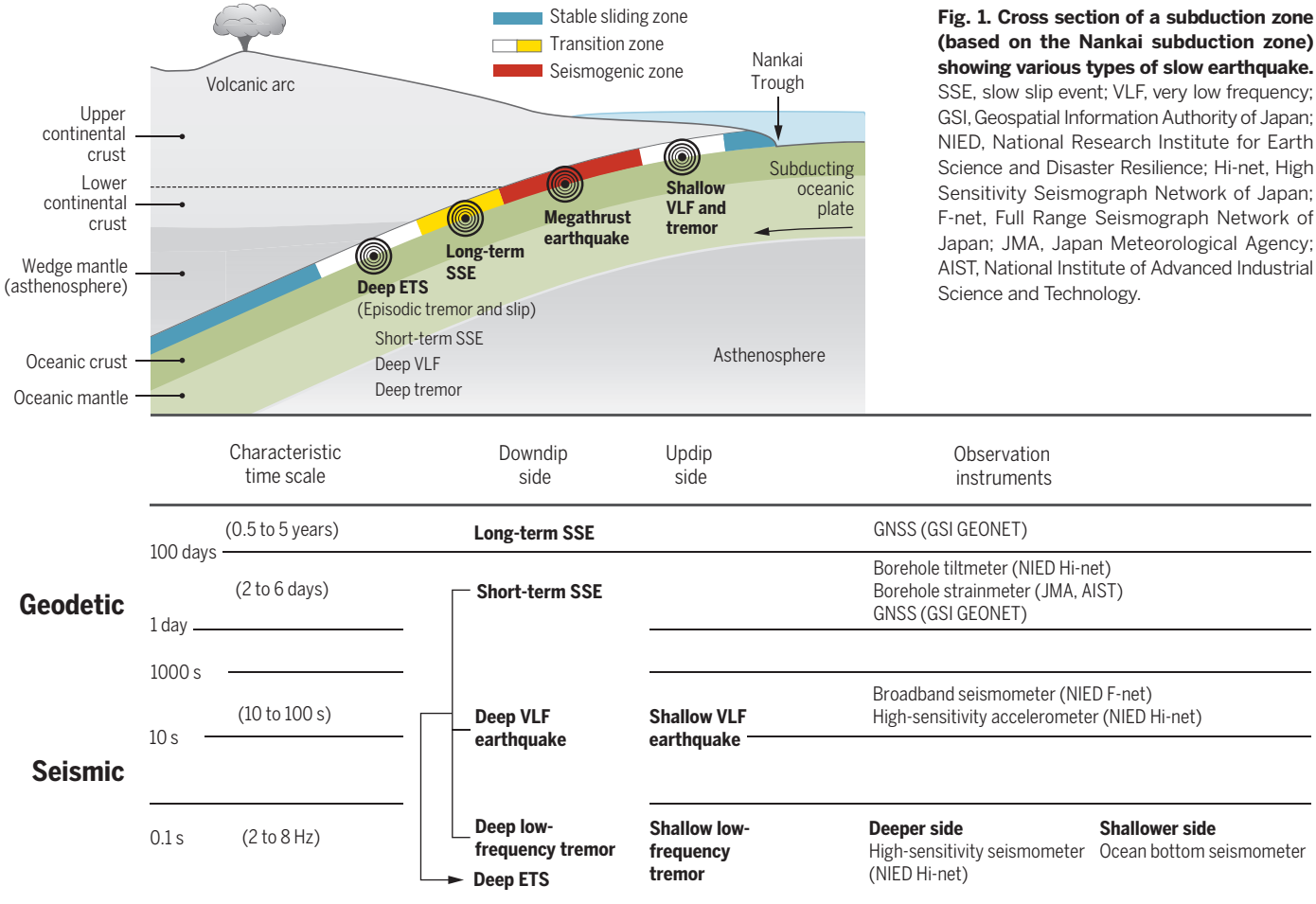



**Fig. 1. Cross section of a subduction zone (based on the Nankai subduction zone) showing various types of slow earthquake.** SSE, slow slip event; VLF, very low frequency; GSI, Geospatial Information Authority of Japan; NIED, National Research Institute for Earth Science and Disaster Resilience; Hi-net, High Sensitivity Seismograph Network of Japan; F-net, Full Range Seismograph Network of Japan; JMA, Japan Meteorological Agency; AIST, National Institute of Advanced Industrial Science and Technology.

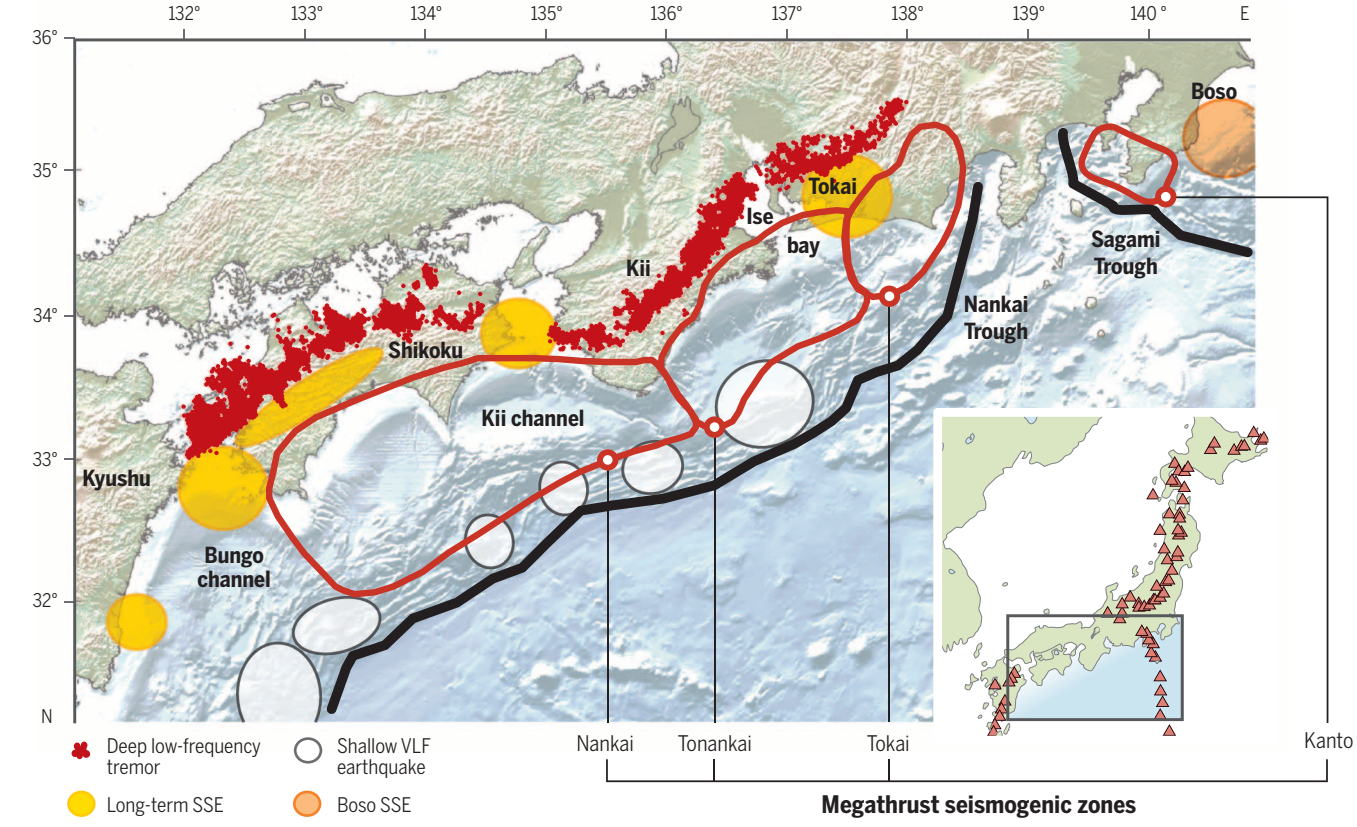



**Fig. 2. Distribution of slow earthquakes in southwest Japan.** Deep low-frequency tremor is denoted by red symbols. Locations where long-term SSEs and shallow VLF earthquakes occur are indicated. The megathrust seismogenic zones along the Nankai subduction zone are outlined in red. The inset map shows the regional setting of the area of interest; active volcanoes are denoted by red triangles.

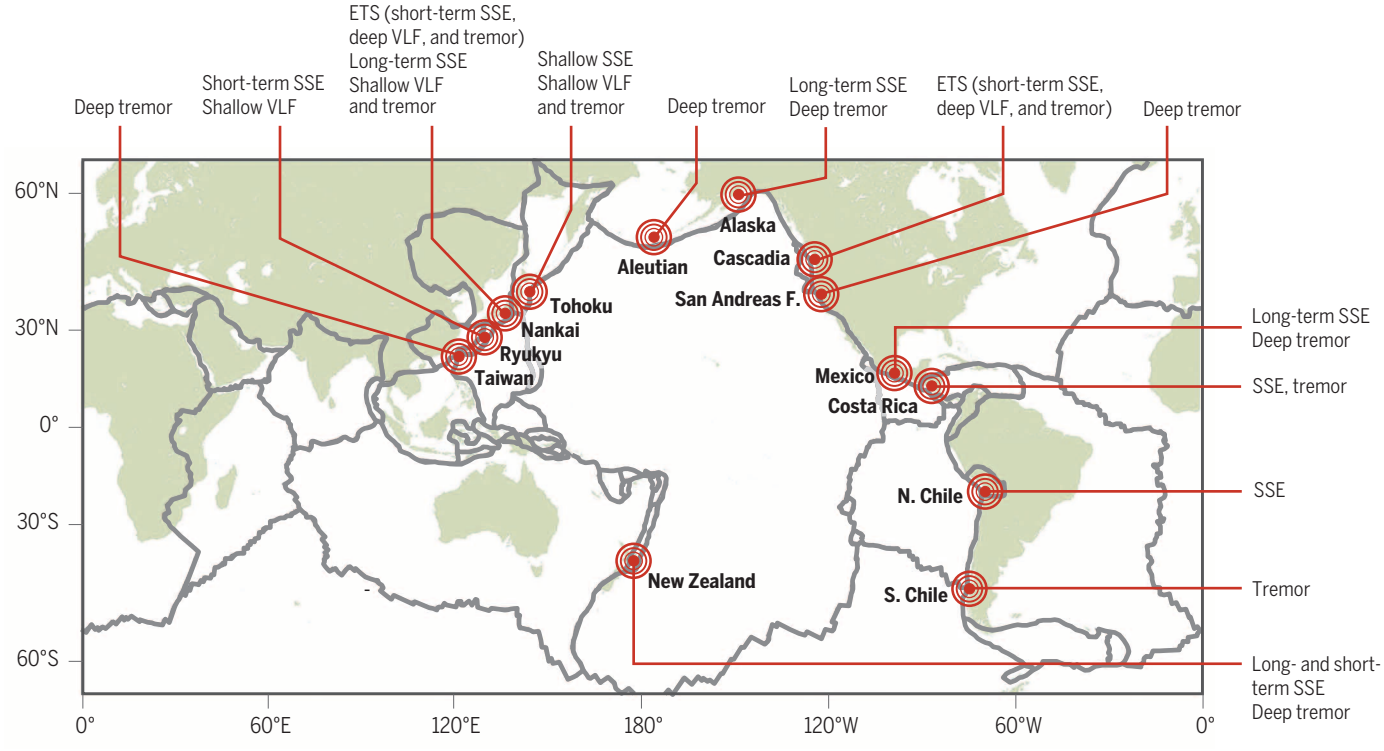



**Fig. 3. Global distribution of slow earthquakes.**

inter-ETS periods, triggered tremor is often associated with the propagation of surface waves from distant, large earthquakes (*38*). The sensitive nature of tremor means that slow earthquakes may play an important role as stress meters, reflecting stress accumulation at strongly locked sections of the megathrust adjacent to slow earthquake regions.

At Parkfield, California, two possible changes in tremor activity along the San Andreas Fault have been recognized prior to the 2004 $M_w$ 6.0 Parkfield earthquake. One is an increase in tremor seismicity beginning approximately 3 weeks before the earthquake (*39*). This foretremor might have been caused by preslip at the mainshock nucleation point, as indicated by an elevation in tremor seismicity that decays gradually after the earthquake, consistent with afterslip (*40*). Another change is a characteristic migration pattern of seismicity revealed by precise event catalogs (*41*). Tremor usually migrates in both directions along the fault; however, during the 3 months before the $M_w$ 6 earthquake, all tremor events migrated unilaterally in a southward direction toward the rupture initiation point of the $M_w$ 6 earthquake. This change in migration direction might indicate stress accumulation around the rupture initiation point.

Relative changes in shear stress at depth can be inferred from temporal changes in the recurrence interval of tremor along the San Andreas Fault. The tremor pattern regularly oscillated at doubled recurrence intervals of 3 and 6 days for many years, but was disrupted by the 2004 Parkfield earthquake (*42*). Period-doubling tremor can be modeled as regular slip in slow and fast ruptures by means of physics-based modeling (*43*). Sudden changes, and gradual recovery, of the recurrence intervals can be explained by temporal changes in shear stress driven by a rise in pore pressure following the Parkfield earthquake.

Slow earthquakes and their relationship to huge earthquakes have been well studied using numerical simulations (*35*, *44*). Matsuzawa *et al.* (*35*) modeled a locked zone with high effective normal stress and a downdip ETS zone with high pore fluid pressure on the dipping interface, adopting a rate- and state-dependent friction law with cutoff velocities. Allowing lateral variations of fluid pressure in the ETS zone, they reproduced the recurrence interval of short- and long-term SSEs and the depth-dependent behavior of ETS. According to their simulation, the recurrence intervals of long- and short-term SSEs gradually become shorter as the huge earthquake approaches. The shortening of recurrence interval is caused by stress accumulation due to a gradual unlocking near the edge of the strongly locked region. After the earthquake, the recurrence intervals return to larger values. This result suggests that shortening of recurrence intervals may indicate an impending huge earthquake. We are unsure whether this result is real, as it may result from the simplicity of the model. However, if the modeling results are correct, they point to the possibility that precursory changes in slow earthquake activity may be indicative of an upcoming huge earthquake. Integrating monitoring with slow earthquake simulation will not only deepen our understanding of this connection, but also potentially allow us to evaluate the timing of events from a long-term perspective. The best-case scenario might be an improvement in the estimation of the probability of earthquake hazard, as slow earthquakes provide a different type of information that is not yet being used.

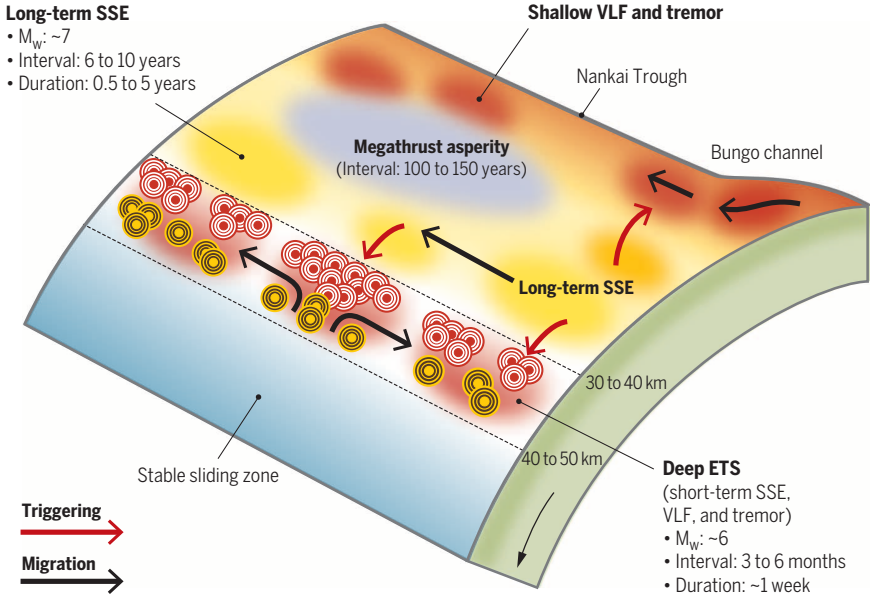



**Fig. 4. Schematic view of the heterogeneous distribution of various types of slow earthquakes in the Nankai subduction zone.** Arrows indicate interactions between the long-term SSEs and other slow earthquakes on the interface of the subducting Philippine Sea plate in southwest Japan. Red and yellow circles indicate tremor at updip and downdip parts within the deep ETS zone, respectively.

One of the striking properties of slow earthquakes is the interaction between different types of such events. These interactions are explained by stress transfer from larger slow earthquakes to triggered slow earthquakes, as discussed below. Differences in the type of slow earthquake activity that is triggered may reflect shear strength or stress level variations therein. During the 2003 and 2010 long-term SSEs in the Bungo channel, shallow VLF earthquakes were triggered (*17*) as well as deep tremor (Fig. 4). The association of shallow VLF earthquakes with shallow tremor activity was also seen in seafloor observations for the 2013 shallow slow earthquake episode (*19*). If we compare the 2010 and 2013 shallow slow earthquake episodes, both are seen to initiate at the southern part of the zone and both migrate in the same direction northward. In 2013, the shallow slow earthquake activity terminated at the point where the Nankai trough bends; however, during the 2010 long-term SSE, the shallow slow earthquake episode propagated around this bend (*45*). We interpret this process of overcoming bends as a weakening of interplate coupling due to the long-term SSE.

## Possible stress transfer from slow to huge earthquakes

Understanding the interactions between huge earthquakes and other phenomena would be very valuable for evaluating the potential of an impending huge earthquake. So far, we have observed several kinds of interaction: (i) between different slow earthquakes; (ii) between slow earthquakes and smaller earthquakes; and (iii) between slow earthquakes and larger earthquakes. The most obvious interaction between slow earthquakes occurred during the long-term SSE in the Bungo channel when, for a few months, minor tremor activity intensified in the updip part of the tremor zone closest to the SSE source fault. However, the rate of deeper tremor activity was constant, regardless of SSE activity (*17*). The tremor response to the SSE clearly depends on distance from the SSE source fault. Therefore, the long-term SSE seems to occur spontaneously and triggers tremor in the adjacent downdip region (Fig. 4). A similar interaction between long-term SSEs and tremor is observed in Mexico. During the 2006 SSE in this region, tremor occurred on the shallower side of the SSE, even though deeper tremor was more persistent (*46*). In contrast, the long-term SSE in Tokai lacks any concurrent, sharp increase in tremor count like that seen in the Bungo channel, even though the relative distance between tremor and SSE source fault is almost the same. The degree of interaction is roughly dependent on the slip velocity of the long-term SSE; slip velocity in the Bungo channel (~3 cm/month) is six times that at Tokai (~0.5 cm/month), resulting in greater production of tremor.

Recently, a series of small long-term SSEs was detected in the gap between the Nankai earthquake rupture area and the tremor zone (*47*). These SSEs appear to migrate eastward along the gap adjacent to the Bungo channel SSE. Tremor

count at the updip part of the deep tremor zone shows a slight increase due to a downward stress loading by these migrating SSEs. The front of the increase in tremor activity propagates quite slowly (~25 km/year). This slow-migrating SSE seems to be a kind of afterslip that is enhanced by the Bungo channel long-term SSE, as afterslip following the 1946 Nankai earthquake occurred in the same region. It is possible that the gap zone is weakly locked, thereby facilitating afterslip in response to neighboring large slip events.

The Boso SSE is a typical example of a large-magnitude SSE driving a swarm of small earthquakes in the downdip region of the major slip patch of the SSE (*23*). The moment magnitude of the SSE and the largest triggered earthquake are generally $M_w$ 6.6 and $M_w$ 5.3, respectively. The duration of the SSE and the triggering of downdip events resemble the behavior observed during the long-term SSE in the Bungo channel. Therefore, the mechanism enabling the Boso SSE to drive earthquake swarms might be similar to that which drives tremor related to the Bungo channel SSE. However, these SSEs differ in their source depth; the Boso SSE source fault is up to 20 km deep, which is shallower than that of the Bungo channel SSE, and so the thermal regime in Boso is colder than that in Bungo. The different thermal regime in the two cases might affect the type of slip seen in the triggered events.

Before the 2011 Tohoku earthquake, with the exception of afterslip, no slow earthquakes had been reported off Tohoku. However, precise studies using data retrieved by on-land networks and offshore seafloor instruments subsequently revealed that slow earthquakes with various time scales in and around the rupture area of the Tohoku earthquake had loaded the mainshock fault. A decade-long SSE took place in the deepest part of the Tohoku mainshock rupture area and accelerated over time (*48*). Near the mainshock rupture initiation point, SSEs were detected geodetically by seafloor pressure gauges in 2008 and 2011 in the vicinity of the high-slip area of the mainshock rupture zone (*49*). During the foreshock sequence lasting 23 days before the Tohoku earthquake, two sequences of SSE migrated toward the mainshock rupture initiation point (*50*). In addition, shallow tremor was detected by seafloor seismometers during the foreshock sequence (*51*). These observations indicate that the plate interface near the mainshock initiation point had been locally unlocked by these slow earthquakes during the foreshock sequence.

Similar slow earthquakes preceding huge earthquakes have been observed in Mexico and Chile. The $M_w$ 7.3 Papanoa (Mexico) earthquake of 18 April 2014 was clearly triggered by an SSE that started in February 2014 in the region neighboring the earthquake source fault (*52*). Before the 2014 $M_w$ 8.2 Iquique earthquake in northern Chile, an SSE lasting about 2 weeks occurred prior to the mainshock. The SSE was located near a strongly locked section of the megathrust, as revealed by both seismicity analysis and geodetic slip inversion of GPS data recorded near the coastline (*53*). The spatiotemporal evolution of intense foreshocks indicates that the SSE migrated toward the rupture initiation point of the mainshock, as also seen before the Tohoku earthquake (*54*, *55*). Of course, accumulated stress on the strongly locked area is required to be close to a critical level for SSEs to prompt unstable dynamic rupture of the mainshock. When an SSE migrated near the strongly locked area, the locked area did not always rupture dynamically, which suggests that forecasts of an impending earthquake cannot rely solely on a migrating pattern of SSEs.

A combination of repeating earthquake analysis and crustal deformation analysis has revealed quasi-periodic slow slip behavior that is widely distributed across the megathrust zone off Tohoku (*56*). The recurrence interval of the slow slip transients ranges from 1 to 6 years and often coincides with or precedes clusters of large interplate earthquakes (i.e., $M > 5$), including the 2011 Tohoku earthquake. These periodic SSEs cause periodic stress loading onto the plate boundary fault and modulate the recurrence time of large earthquakes.

## Conclusion

Slow earthquakes have been frequently detected in regions neighboring megathrust seismogenic zones. Some types of slow earthquakes have occurrence styles somewhat similar to those of megathrust earthquakes. The study of slow earthquakes as an analog to megathrust earthquakes is expected to improve our understanding of the megathrust rupture cycle. Furthermore, a better understanding of the various slow earthquake types may deepen our knowledge of the tectonic framework required for the formation of subduction systems, including the rupture style of huge megathrust earthquakes. Slow earthquakes may be able to play a role as stress meters or in stress transfer. Frequent loading might eventually trigger failure of the megathrust, although the stress transfer attributable to each slow earthquake episode will be very small. However, many slow earthquakes do not lead directly to megathrust earthquakes, because the final rupture of the megathrust earthquake depends on the areal extent of the megathrust fault that is close to failure and how close the critical area is to failure in the seismogenic zone (*57*). The most important and useful information for evaluating seismic potential would be an estimate of the degree of criticality within the seismogenic zone adjacent to slow earthquake source region. Long-term monitoring of slow earthquakes is required so that a reliable picture of these phenomena can be built over all time scales, and so that physics-based numerical simulations that reproduce the observed plate boundary faulting behavior can be developed.

### REFERENCES AND NOTES


1. K. Nanjo, N. Hirata, K. Obara, K. Kasahara, *Geophys. Res. Lett.* **39**, L20304 (2012).
2. S. Tanaka, *Geophys. Res. Lett.* **39**, L00G26 (2012).
3. T. Tormann, B. Enescu, J. Woessner, S. Wiemer, *Nat. Geosci.* **8**, 152–158 (2015).
4. W. Wang, P. M. Shearer, *J. Geophys. Res.* **120**, 6317–6328 (2015).
5. T. Sagiya, *Earth Planets Space* **56**, xxix–xli (2004).
6. Y. Okada *et al.*, *Earth Planets Space* **56**, xv–xxviii (2004).
7. R. D. Hyndman, K. Wang, M. Yamano, *J. Geophys. Res.* **100**, 15373–15392 (1995).
8. Z. Peng, J. Gomberg, *Nat. Geosci.* **3**, 599–607 (2010).
9. S. Mazzotti, J. Adams, *Bull. Seismol. Soc. Am.* **94**, 1954–1959 (2004).
10. K. Obara, *Science* **296**, 1679–1681 (2002).
11. K. Obara, H. Hirose, F. Yamamizu, K. Kasahara, *Geophys. Res. Lett.* **31**, L23602 (2004).
12. Y. Ito, K. Obara, K. Shiomi, S. Sekine, H. Hirose, *Science* **315**, 503–506 (2007).
13. G. Dragert, K. Wang, T. S. James, *Science* **292**, 1525–1528 (2001).
14. G. Rogers, H. Dragert, *Science* **300**, 1942–1943 (2003).
15. A. Kobayashi, *Earth Planets Space* **66**, 9 (2014).
16. H. Suito, S. Ozawa, *J. Seismol. Soc. Jpn.* **61**, 113–135 (2009) (in Japanese).
17. H. Hirose *et al.*, *Science* **330**, 1502 (2010).
18. K. Obara, Y. Ito, *Earth Planets Space* **57**, 321–326 (2005).
19. Y. Yamashita *et al.*, *Science* **348**, 676–679 (2015).
20. M. Nakamura, N. Sunagawa, *Geophys. Res. Lett.* **42**, 1076–1082 (2015).
21. T. Nishimura, *Prog. Earth Planet. Sci.* **1**, 22 (2014).
22. Y. Asano, K. Obara, Y. Ito, *Earth Planets Space* **60**, 871–875 (2008).
23. H. Hirose, H. Kimura, B. Enescu, S. Aoi, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 15157–15161 (2012).
24. M. R. Brudzinski, R. M. Allen, *Geology* **35**, 907–910 (2007).
25. A. Ghosh, E. Huesca-Pérez, E. Brodsky, Y. Ito, *Geophys. Res. Lett.* **42**, 3228–3232 (2015).
26. J. S. Payero *et al.*, *Geophys. Res. Lett.* **35**, L07305 (2008).
27. L. M. Wallace, J. Beavan, *J. Geophys. Res.* **115**, B12402 (2010).
28. L. M. Wallace *et al.*, *Science* **352**, 701–704 (2016).
29. R. M. Nadeau, D. Dolenc, *Science* **307**, 389 (2005).
30. Z. Peng, K. Chao, *Geophys. J. Int.* **175**, 825–829 (2008).
31. K. Satake, *Geosci. Lett.* **1**, 15 (2014). 10.1186/s40562-014-0015-7
32. A. G. Wech, N. M. Bartlow, *Geophys. Res. Lett.* **41**, 392–398 (2014).
33. S. Ozawa, H. Yarai, T. Imakiire, M. Tobita, *Earth Planets Space* **65**, 67–73 (2013).
34. S. Ozawa, *Geophys. Res. Lett.* **41**, 2762–2768 (2014).
35. T. Matsuzawa, H. Hirose, B. Shibazaki, K. Obara, *J. Geophys. Res.* **115**, B12301 (2010).
36. A. Kato, T. Igarashi, K. Obara, *Geophys. Res. Lett.* **41**, 5868–5874 (2014).
37. J. L. Rubinstein, M. La Rocca, J. E. Vidale, K. C. Creager, A. G. Wech, *Science* **319**, 186–189 (2008).
38. K. Chao *et al.*, *Bull. Seismol. Soc. Am.* **103**, 1551–1571 (2013).
39. R. M. Nadeau, A. Guilhem, *Science* **325**, 191–193 (2009).
40. F. Brenguier *et al.*, *Science* **321**, 1478–1481 (2008).
41. D. R. Shelly, *Geophys. Res. Lett.* **36**, L17318 (2009).
42. D. R. Shelly, *Science* **328**, 1385–1388 (2010).
43. D. M. Veedu, S. Barbot, *Nature* **532**, 361–365 (2016).
44. K. Ariyoshi *et al.*, *Earth Planets Space* **64**, 693–702 (2012).
45. Y. Asano, K. Obara, T. Matsuzawa, H. Hirose, Y. Ito, *Geophys. Res. Lett.* **42**, 331–338 (2015).
46. W. B. Frank *et al.*, *J. Geophys. Res.* **119**, 7686–7700 (2014).
47. R. Takagi, K. Obara, T. Maeda, *Geophys. Res. Lett.* **43**, 1066–1074 (2016).
48. Y. Yokota, K. Koketsu, *Nat. Commun.* **6**, 5934 (2015).
49. Y. Ito *et al.*, *Tectonophysics* **600**, 14–26 (2013).
50. A. Kato *et al.*, *Science* **335**, 705–708 (2012).
51. Y. Ito, R. Hino, S. Suzuki, Y. Kaneda, *Geophys. Res. Lett.* **42**, 1725–1731 (2015).
52. M. Radiguet *et al.*, paper presented at the American Geophysical Union Fall Meeting, San Francisco, 18 December 2015.
53. S. Ruiz *et al.*, *Science* **345**, 1165–1169 (2014).
54. A. Kato, S. Nakagawa, *Geophys. Res. Lett.* **41**, 5420–5427 (2014).
55. A. Kato, J. Fukuda, T. Kumazawa, S. Nakagawa, *Sci. Rep.* **6**, 24792 (2016).
56. N. Uchida, T. Iinuma, R. M. Nadeau, R. Bürgmann, R. Hino, *Science* **351**, 488–492 (2016).
57. H. Kanamori, E. E. Brodsky, *Phys. Today* **54**, 34–40 (2001).


10.1126/science.aaf1512

RESEARCH ARTICLE SUMMARY

MITOCHONDRIA

# ER-mitochondria contacts couple mtDNA synthesis with mitochondrial division in human cells

Samantha C. Lewis, Lauren F. Uchiyama, Jodi Nunnari*

**INTRODUCTION:** Mitochondria are endosymbiotic organelles that have their own genome and perform many essential functions in eukaryotic cells, including ATP synthesis via oxidative phosphorylation. Mitochondria evolved from bacteria, which stringently replicate and segregate their genome in a binary fission process. The residual circular ~16-kilobase human mitochondrial genome is essential, as it encodes mitochondrial ribosomal and transfer RNAs and respiratory chain complex proteins. Within human cells, hundreds to thousands of copies of mitochondrial DNA (mtDNA) are packaged into nucleoids, the unit of mtDNA inheritance, and distributed within dynamic mitochondrial networks. The nature of mtDNA transmission in mitochondrial syncytia is relaxed, with replication occurring asynchronously throughout the cell cycle and also in postmitotic cells. If and how nucleoids are actively chosen for mtDNA replication and distributed within mitochondrial networks is not understood. This question is highly relevant for understanding the basis of human metabolic diseases caused by mutations in mtDNA and in nuclear genes that affect mtDNA maintenance. In addition, aging and neurodegenerative disorders are also linked to defective mtDNA maintenance and mitochondrial dysfunction. In this study, we investigated the fundamental process of mtDNA transmission in mammalian cells.

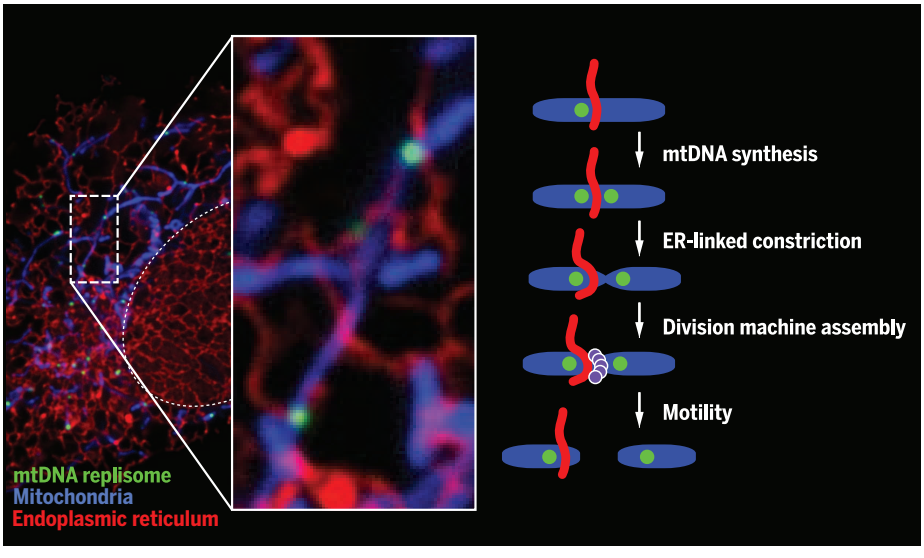
ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aaf5549

**RATIONALE:** To address the cellular mechanism for mtDNA transmission, we examined whether nucleoid and mitochondrial distribution were coupled. Mitochondrial distribution is determined by mitochondrial division, fusion, and motility events. Previous work established that the mitochondrial division site placement is not random and is instead spatially marked by regions of contact between the endoplasmic reticulum (ER) and mitochondria, in a process termed ER-associated mitochondrial division (ERMD). In biological systems, division events serve a fundamental role in the inheritance of genetic material. Thus, we addressed whether ERMD serves to facilitate the transmission of mtDNA by examining the behavior of mitochondria, ER, and nucleoids in mammalian cells via fluorescent microscopy. We marked the subset of nucleoids in cells actively engaged in mtDNA synthesis with a functional green fluorescent protein–tagged version of POLG2, the processivity subunit of the human mitochondrial DNA polymerase holoenzyme. Using this sensitive and highly specific marker for mtDNA synthesis in live cells, we asked whether replicating nucleoids were selectively linked to ERMD for the purpose of ensuring the segregation of nascent mtDNA to daughter mitochondria.

**RESULTS:** Our work revealed that nucleoids actively engaged in mtDNA synthesis in mammalian cells were spatially and temporally linked to a small subset of ER-mitochondria contacts destined for mitochondrial division. At division sites, mtDNA replication occurred upstream of mitochondrial constriction and assembly of the division machinery. Nucleoids containing nascent mtDNA localized to mitochondrial tips, and these products of division were preferentially distributed within cells as compared with nonreplicative nucleoids. Our observations also demonstrated that ER structure and mtDNA maintenance were intertwined; ER tubules proximal to nucleoids were necessary but not sufficient for mtDNA synthesis and also functioned in nucleoid distribution.

**CONCLUSION:** We propose that, at ER-mitochondria contacts destined for division, the consecutive events of mtDNA replication, mitochondrial division, and mitochondrial motility are connected together to ensure the accurate distribution of nucleoids within cells. Our findings suggest that ER-mitochondria contacts coordinate the licensing of mtDNA replication with downstream mitochondrial division events to distribute newly replicated mtDNA to daughter mitochondria. The connection revealed between ER structure and mtDNA replication and distribution has broad implications for understanding human cellular homeostasis and the cellular pathology underlying human diseases. ■




**ER-mitochondria contacts coordinate mtDNA replication with mitochondrial division.** In human cells, a subset of ER-mitochondria contacts are spatially linked to mitochondrial nucleoids engaged in replication and are destined for mitochondrial division. (Left) Light image is of an osteosarcoma U2OS cell; (right) in the schematic depiction, colors are as on the labels to the left; and the replicating nucleoid is marked by POLG2 in green.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: jmnunnari@ucdavis.edu
Cite this article as S. C. Lewis *et al.*, *Science* 353, aaf5549 (2016). DOI: 10.1126/science.aaf5549

RESEARCH ARTICLE

MITOCHONDRIA

# ER-mitochondria contacts couple mtDNA synthesis with mitochondrial division in human cells

Samantha C. Lewis, Lauren F. Uchiyama, Jodi Nunnari*

**Mitochondrial DNA (mtDNA) encodes RNAs and proteins critical for cell function. In human cells, hundreds to thousands of mtDNA copies are replicated asynchronously, packaged into protein-DNA nucleoids, and distributed within a dynamic mitochondrial network. The mechanisms that govern how nucleoids are chosen for replication and distribution are not understood. Mitochondrial distribution depends on division, which occurs at endoplasmic reticulum (ER)–mitochondria contact sites. These sites were spatially linked to a subset of nucleoids selectively marked by mtDNA polymerase and engaged in mtDNA synthesis—events that occurred upstream of mitochondrial constriction and division machine assembly. Our data suggest that ER tubules proximal to nucleoids are necessary but not sufficient for mtDNA synthesis. Thus, ER-mitochondria contacts coordinate licensing of mtDNA synthesis with division to distribute newly replicated nucleoids to daughter mitochondria.**

Mutations in mitochondrial DNA (mtDNA) and in nuclear genes that control mtDNA maintenance cause mitochondrial dysfunction and are linked to human disease and aging, which affirms the functional importance of mtDNA (*1*–*4*). The units of mitochondrial inheritance are mtDNA-protein complexes called nucleoids (*5*). Accurate maintenance of mtDNA requires replication, repair, packaging, and distribution of the mitochondrial nucleoid at the cellular level. In mammalian cells, mitochondrial DNA replication is mediated by a nuclear-encoded replisome composed of a polymerase gamma holoenzyme containing a catalytic subunit (POLG1) and a processivity subunit (POLG2) (*6*–*10*). In addition to the polymerase complex, the replisome contains the helicase Twinkle and a mitochondria-specific single-stranded DNA binding protein, which together facilitate the formation of a single-stranded DNA replication template (*11*, *12*). Within cells, mtDNA is packaged into a nucleoid by TFAM, a nuclear-encoded DNA-bending protein, which also plays a role in mtDNA replication and transcription (*13*–*18*).

Although molecular players involved in mtDNA replication and packaging have been described, the mechanisms underlying the spatial regulation of mtDNA replication and intracellular nucleoid distribution have been elusive. This is partly because mtDNA is present in cells in multiple copies, and the spatiotemporal regulation of its replication and distribution is relaxed in comparison with the nuclear genome (*19*). Indeed, mtDNA replication occurs asynchronously with the cell cycle and within postmitotic tissues, such as the brain and muscle (*20*, *21*). Nucleoids are evenly distributed within mitochondria and constrained in their motility (*5*, *22*). Mitochondrial distribution is in large part dependent on cytoskeletal-based motility and on mitochondrial division (*23*), mediated in mammalian cells by DRP1, a cytosolic dynamin-related guanosine triphosphatase (GTPase) that forms assemblies around mitochondria to facilitate membrane scission (*24*, *25*). DRP1 recruitment and assembly occur at sites of endoplasmic reticulum (ER)–mitochondrial contact, where mitochondrial constriction is also observed (*26*). Perturbation of mitochondrial division in both yeast and mammalian cells causes nucleoid aggregation, mtDNA deletions, and mtDNA depletion, which suggests a fundamental and functional link between mitochondrial division and mtDNA maintenance (*27*–*29*). In yeast, ER-linked division sites, marked by the fungus-specific ER-mitochondria encounter structure (ERMES) complex, are spatially linked to nucleoids, which further suggests a role for ER-mitochondria contacts in mtDNA maintenance (*30*). Here, we asked whether ER-mitochondria contact sites function to couple mtDNA replication with mitochondrial division for the purpose of distributing newly replicated mtDNA in human cells.

## ER-mitochondria contacts and ER-associated mitochondrial division are spatially linked to nucleoids in mammalian cells

We asked if ER-associated mitochondrial division (ERMD) events are spatially linked to mitochondrial nucleoids in human cells by simultaneously imaging mitochondria, nucleoids, and the ER network at high spatial and temporal resolution using spinning disk confocal microscopy. Osteosarcoma cells (U2OS) were transiently transfected with green fluorescent protein (GFP)–tagged TFAM (TFAM-GFP), a well-characterized marker of the total nucleoid population, mitochondrial matrix–targeted blue fluorescent protein (mito-BFP), and ER-targeted mCherry or mRuby (Sec61b-mCherry or mRuby-KDEL). TFAM-GFP–labeled foci were evenly spaced within mitochondria, as previously described for nucleoids in other cell types (Fig. 1A) (*5*, *16*, *18*). TFAM-GFP–labeled nucleoids were also localized adjacent to points where ER tubules crossed over mitochondria in a perpendicular fashion (Fig. 1A, right), and a subset of nucleoids remained stably linked to ER-mitochondria contacts over time, despite ER network remodeling and mitochondrial motility (Fig. 1B, arrowheads). We further assessed the spatial link between ER-mitochondria contacts and nucleoids by determining the Pearson correlation coefficient of mRuby-KDEL and TFAM-GFP fluorescence intensity along line scans of mitochondria imaged in live U2OS cells ($n$ = 58). Consistent with our observations, this analysis indicated a highly significant enrichment of ER signal specifically within 17 pixels (~1 μm) laterally adjacent to nucleoids (Pearsons' $R$ = 0.59) (fig. S1A). Thus, in general, nucleoids are spatially linked to ER-mitochondria contacts in human cells.

Accordingly, we observed that a majority of ERMD events (82%, $n$ = 62) were spatially linked to nucleoids (within a 1-μm distance), which resulted in their localization at mitochondrial tips after division (Fig. 1, C and arrowhead in D). We also performed time-lapse imaging of nucleoids and mitochondria in COS-7 primate cells labeled with the selective vital dyes PicoGreen DNA stain and MitoTracker Red, respectively. Retrospective nucleoid tracking over time revealed that nucleoids at mitochondrial tips had been displaced a greater distance on average (3 times as great over a time period of 12.5 min) than intramitochondrial nucleoids, which suggested that the subset of nucleoids linked to ERMD sites are preferentially distributed within cells (Fig. 1E).

## ER-mitochondria contacts are not rate-limiting for mitochondrial division

The high density of ER and mitochondrial networks in mammalian cells suggested that ER-mitochondria contact sites may not be rate-limiting for ERMD. To test this, we defined and quantified persistent ER-mitochondria contacts in cells by time-lapse imaging of mitochondria and ER for 5 min at 15-s intervals using mito-BFP and mRuby-KDEL/Sec61b-mCherry, respectively, in U2OS cells. ER and mitochondria were segmented in time-lapse images by thresholding, and regions of overlap between the organelles were identified and tracked over 5 min (see fig. S1B, arrowheads). Most regions of ER-mitochondria colocalization were transient, but ~100 distinct regions per cell were identified in which the ER and mitochondria persistently colocalized over the duration of imaging (fig. S1C). We followed the fate of persistent ER-mitochondria colocalized regions in relation to mitochondrial division over the duration of imaging. At a small fraction

Department of Molecular and Cellular Biology, University of California, Davis, CA 95616, USA.
*Corresponding author. Email: jmnunnari@ucdavis.edu

of these regions, mitochondrial constrictions were observed and/or developed (arrowheads in Fig. 1F; also Fig. 1G). An even smaller fraction of ER-mitochondria colocalized regions was linked to division events, consistent with published observations indicating that ER-linked mitochondrial constriction precedes division (Fig. 1G) (*26*). Thus, although there are many ER-mitochondria contacts, only a subset are destined to be linked to ERMD. Given that ER-mitochondria contacts are also linked spatially to nucleoids in general (fig. S1A), we considered whether a functionally specialized subset of nucleoids marks ERMD events in cells.

### Nucleoids engaged in mtDNA synthesis mark nascent mitochondrial division sites at ER-mitochondria contacts

It has been proposed that functional subpopulations of nucleoids coexist in mammalian cells, distinguished on the basis of mtDNA replication and/or transcription status (*31*, *32*). In

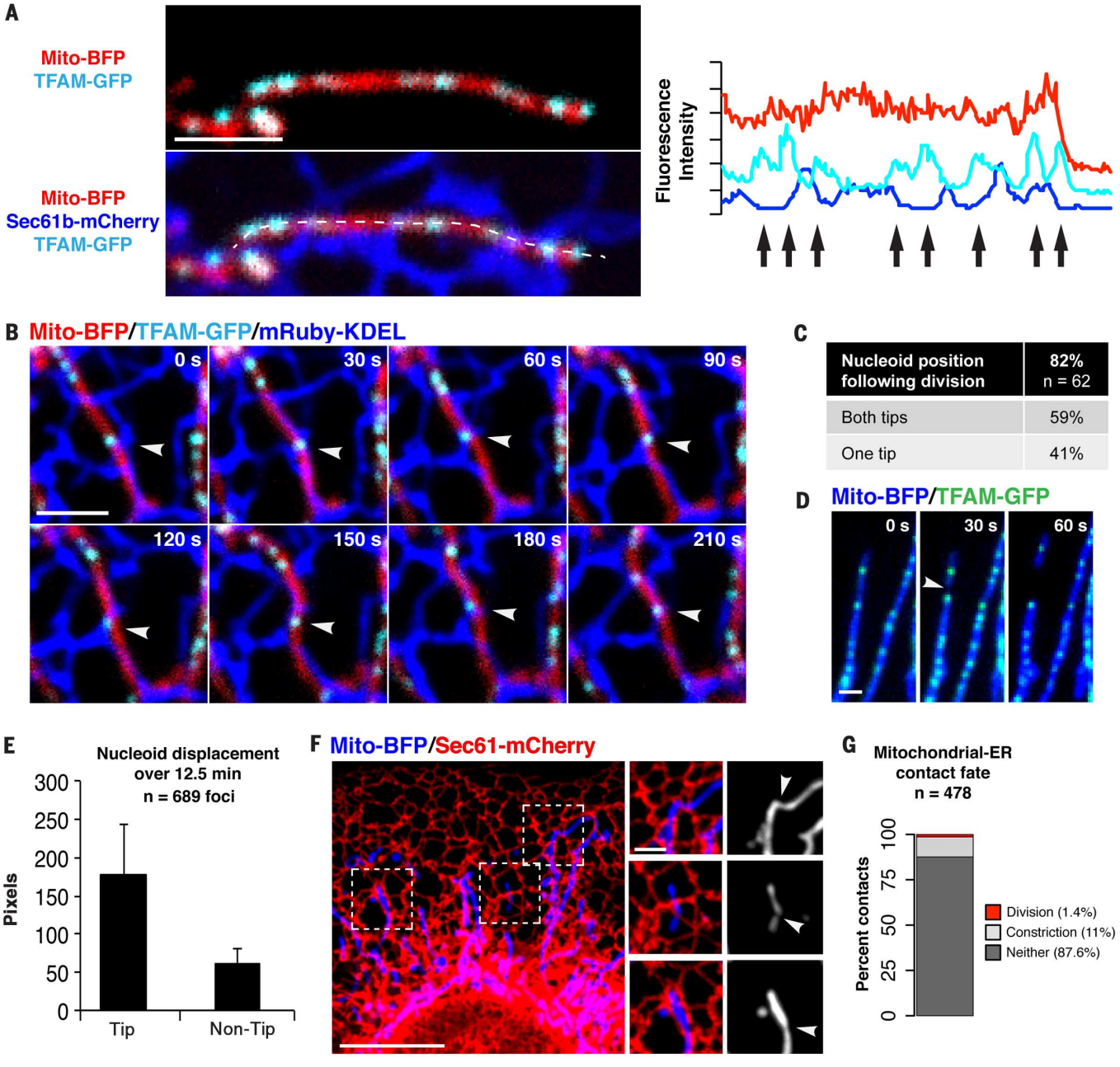


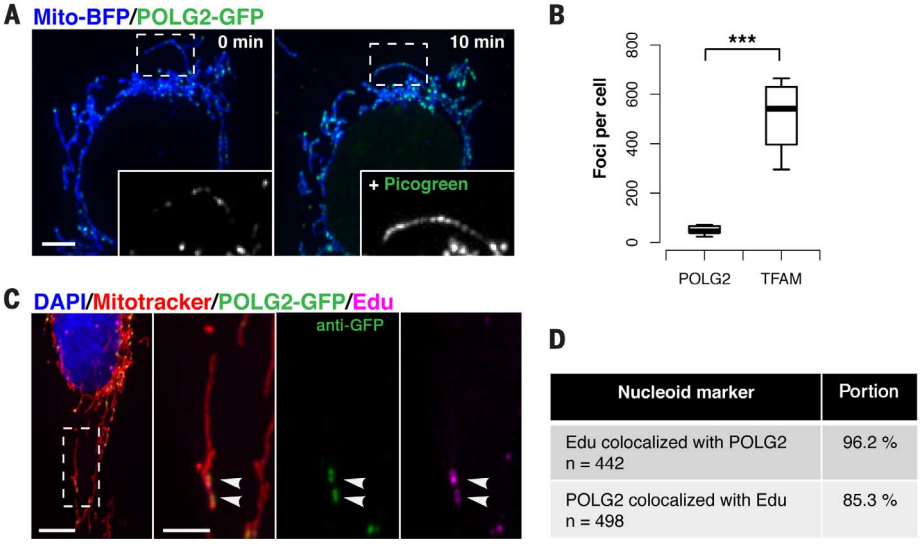

**Fig. 1. Mitochondrial DNA nucleoids are spatially linked to mitochondria-ER contacts in human cells.** (**A**) (Left) Panels show a merged image of a live U2OS cell expressing mito-BFP, TFAM-GFP, and Sec61b-mCherry (ER). (Right) The pixel intensity of mito-BFP, TFAM-GFP, and Sec61b-mCherry from a line scan drawn along the mitochondrial tubule (dashed line), arrows indicate nucleoid positions. (**B**) Time-lapse images of a U2OS cell expressing mito-BFP, TFAM-GFP, and mRuby-KDEL (ER); a single plane is shown. Arrowheads indicate a site of persistent colocalization between a TFAM-GFP–labeled nucleoid and an ER-mitochondria contact. (**C**) Number of mitochondrial divisions in U2OS cells spatially linked to TFAM-GFP–labeled nucleoids, from 43 cells. (**D**) Time-lapse images of mitochondrial division (marked by arrowhead) spatially linked to a TFAM-labeled nucleoid focus in a U2OS cell. (**E**) The displacement of PicoGreen-labeled nucleoids in live COS-7 cells over 12.5 min as a function of their intramitochondrial position. Data are means ± SD. (**F**) (Left) Merged image of a live U2OS cell expressing mito-BFP and Sec61b-mCherry (ER). (Right) Examples of mitochondrial constrictions colocalized with ER tubules (arrowheads). (**G**) The percentage of persistent mitochondrial-ER colocalizations that become sites of mitochondrial constriction or division over 5 min in live U2OS cells. Scale bars: (A), (B), (D), and inset in (F), 2 μm; (F), 10 μm.

yeast, a subpopulation of nucleoids is spatially linked to ERMES foci, which in turn mark a fraction of mitochondrial division sites (*30*, *33*). Thus, we tested whether in mammalian cells mtDNA synthesis is specifically coupled to ERMD, which might help to ensure the segregation of nascent mtDNA to daughter mitochondria.

To visualize nucleoids engaged in mtDNA synthesis in cells, we used POLG2-GFP, a fluorescently tagged version of the processivity subunit of the human mitochondrial DNA polymerase holoenzyme, whose enzymatic properties and affinity for mtDNA are indistinguishable from untagged POLG2 in vitro (*34*). Sixteen hours after transfection of live U2OS cells, POLG2-GFP labeled a subset of the total nucleoid population within mitochondria, which was revealed by the addition of PicoGreen DNA stain and subsequent reimaging (Fig. 2A and fig. S2A). Consistent with this, comparison of the total number of nucleoids labeled with TFAM-GFP to the number of POLG2-GFP–labeled nucleoids in live U2OS cells indicated that there were significantly fewer POLG2-GFP foci per cell (9.4% of total nucleoids) (Fig. 2B). To test whether the POLG2-labeled nucleoids were selectively engaged in mtDNA synthesis, live cells expressing POLG2-GFP and labeled with MitoTracker Red, were incubated with the nucleotide thymidine analog 5-ethynyl-2-deoxyuridine (EdU) for 1 hour, which was subsequently visualized in fixed cells with AlexaFluor647 using copper click chemistry (*35*). The vast majority of POLG2-GFP foci (as detected via α-GFP–AlexaFluor488) colocalized with detectable EdU incorporation at nucleoids within mitochondria (96% from 15 cells) (Fig. 2, C and D). Consistently, there was no significant difference between the number of EdU or POLG2-GFP foci per cell in either live or fixed cells (fig. S2B). We also assessed whether expression of POLG2-GFP perturbed mtDNA maintenance by examining mtDNA copy number by quantitative polymerase chain reaction (qPCR). We detected no difference in mtDNA levels between cells expressing POLG2-GFP versus mock-transfected cells (fig. S2C). In contrast, and as previously shown, a significant increase in mtDNA copy number was observed in cells overexpressing TFAM-GFP, which indicated that U2OS cells were capable of modulating mtDNA synthesis during our experiments (*13*, *36*). Thus, exogenously expressed POLG2-GFP is recruited to the endogenous mtDNA replisome and provides a sensitive and highly specific marker for monitoring mtDNA synthesis in live cells.

With this tool in hand, we asked whether nucleoids engaged in mtDNA synthesis, labeled faithfully by POLG2-GFP, were spatially linked to ERMD in a selective manner. As shown in a representative time-lapse series (Fig. 3A), POLG2-GFP–labeled nucleoids were indeed spatially linked to mitochondrial division at ER-mitochondria contact sites in live U2OS cells, as labeled by mRuby-KDEL and mito-BFP markers, respectively. A majority of ERMD events were linked to POLG2-labeled nucleoids (within 1 μm) and ERMD occurred at a rate more than 3 times that expected from random chance, at persistent ER-mitochondria contacts (Fig. 3B). In addition, a comparable proportion of ERMD events were linked to nucleoids in cells expressing either the general nucleoid marker TFAM-GFP or the replication-specific marker POLG2-GFP (82 versus 73%, $P = 0.71$) (Figs. 1C and 3B). Thus, nucleoid-linked ERMD events seem to occur predominantly at nucleoids engaged in mtDNA synthesis.

To further test the idea that nucleoid-linked ERMD events occur predominantly at nucleoids actively engaged in mtDNA synthesis, we exploited the observation that POLG2-GFP nucleoids localized at the tips of daughter mitochondria as a consequence of ERMD (Fig. 3A). Steady-state analysis of nucleoid position in U2OS cells revealed a highly significant spatial enrichment of POLG2-GFP labeled nucleoids within 1 μm of mitochondrial tips, as compared to the total TFAM-GFP labeled nucleoid population (Fig. 4, A and B), despite the nearly 10 times as great number of TFAM-GFP–labeled nucleoids detected per cell (Fig. 2B). To validate that our observations were generally reflective of mammalian mtDNA segregation, we analyzed two additional cell lines: noncancerous ARPE19 retinal epithelial cells and COS-7 primate fibroblasts. As in U2OS cells, POLG2-GFP–labeled nucleoids were enriched at mitochondrial tips in these additional lines, where they were also colocalized with focal EdU labeling in fixed cells (fig. S3, A and B). In addition, we reasoned that if mtDNA replication was selectively linked to ERMD, the number of POLG2-GFP– and/or EdU-labeled nucleoids per mitochondrion would be constrained to the number of mitochondrial tips created by mitochondrial division, as opposed to scaling to the total length of the organelle. Thus, we examined the distribution of replicating and total nucleoids per mitochondrion in live U2OS cells (labeled with TFAM-GFP or POLG2-GFP) and in fixed ARPE19 cells (labeled with PicoGreen and EdU) (Fig. 4C, left and right, respectively). Indeed, consistent with the relatively uniform distribution of nucleoids within mitochondria (Fig. 1A), individual organelles contained numerous TFAM-GFP– or PicoGreen-labeled nucleoids, and the total number of nucleoids per mitochondrion was highly correlated with the length of the organelle (Fig. 4C). In contrast, the number of POLG2-GFP– or EdU-labeled nucleoids per organelle was not well correlated with mitochondrial length, and in a majority of instances, there was a clear constraint on the number of EdU- or POLG2-GFP–labeled nucleoids to two or less per organelle, regardless of organelle length (Fig. 4C). Closer examination of outlier EdU- or POLG2-labeled nucleoids in mitochondria that contained greater than the median number of nucleoids revealed that these organelles were either branched and, consequently, had a greater number of tips with associated EdU- or POLG2-labeled nucleoids, or were unbranched and contained an additional internally localized pair of EdU foci (fig. S4). Internal EdU foci pairs were closely spaced, suggestive of a segregation intermediate. Thus, in mammalian cells, a majority




| Nucleoid marker | Portion |
|---|---|
| Edu colocalized with POLG2 n = 442 | 96.2 % |
| POLG2 colocalized with Edu n = 498 | 85.3 % |

**Fig. 2. POLG2-GFP is specifically recruited to replicating nucleoids in live cells.** (**A**) A live U2OS cell expressing mito-BFP and POLG2-GFP was imaged (left), then stained with PicoGreen DNA dye and reimaged 10 min later (right). (Insets) 488-channel signal intensity in an example mitochondrion (left), and the same organelle after PicoGreen staining (right). Magnified 2×. (**B**) The number of mitochondrial POLG2-GFP ($n$ = total 441 foci from 10 cells) or TFAM-GFP foci ($n$ = 5182 foci from 10 cells) per U2OS cell (***$P < 0.001$, two-tailed $t$ test). Data are means ± SD. (**C**) Representative image of a U2OS cell expressing POLG2-GFP and pulse-labeled with 50 μM EdU, fixed, and stained with 4′,6′-diamidino-2-phenylindole (DAPI) (DNA, blue); MitoTracker (mitochondria, red); anti-GFP–AlexaFluor488 conjugate antibody (POLG2-GFP, green); and Click-iT EdU-AlexaFluor647 (nascent DNA, magenta). Arrowheads indicate colocalization. (**D**) Observed colocalization between mitochondrial POLG2-GFP and EdU foci in fixed U2OS cells labeled as in (C), from 15 cells. Scale bars: (A) and (C), 10 μm; inset in (C), 5 μm.

of ERMD events are spatially linked to the subset of nucleoids that are actively engaged in mtDNA replication, consistent with a role for ERMD in the coordinated segregation of nascent mitochondrial genomes.

To rigorously test this model and to determine the fate of replicating nucleoids linked to ERMD, we performed an EdU pulse-chase analysis of mtDNA in cells under native conditions, in the absence of POLG2-GFP expression (Fig. 4D, top). We pulse-labeled ARPE19 cells with EdU for 1 hour, under conditions where the replication of nuclear DNA was inhibited; chased for 1, 24, or 48 hours; and subsequently analyzed the position of EdU-labeled nucleoids relative to mitochondrial tips after fixation. Consistent with our previous observations (Figs. 3, A and B, and 4, A and B), EdU nucleoids detected during the pulse and after the 1-hour chase were highly enriched within 1 μm of mitochondrial tips (Fig. 4D and fig. S5, A to C, representative images). Quantification revealed that the positioning of EdU-labeled nucleoids relative to tips progressively decreased after the 24- and 48-hour chases, toward a random distribution (Fig. 4D). Thus, replicating nucleoids are indeed both spatially and temporally linked to mitochondrial division, and after the completion of mtDNA replication, nascent mtDNA is segregated into daughter mitochondria and subsequently distributed away from the preceding division sites.

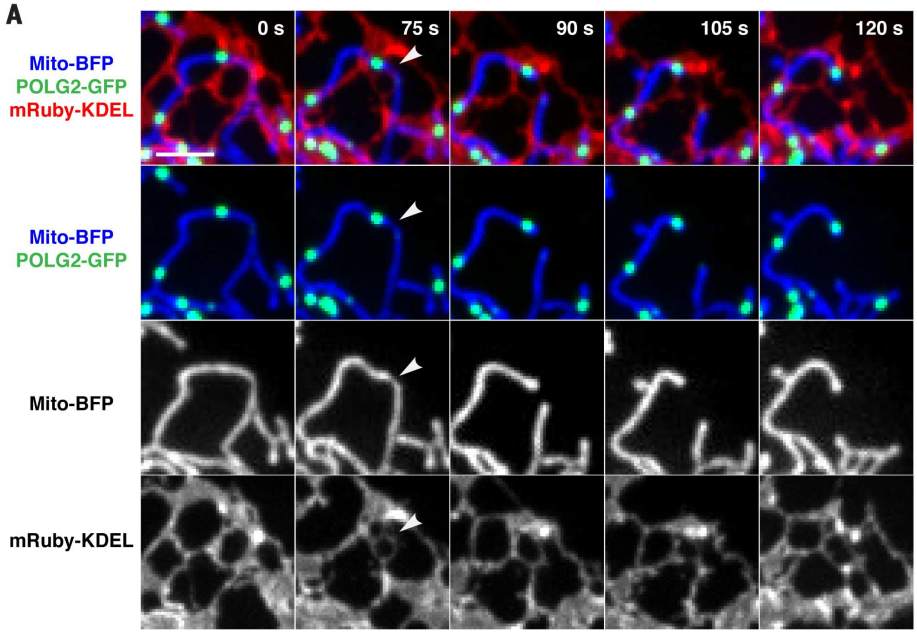
### The mtDNA replisome is an early marker of nascent mitochondrial division sites

To gain further insight into the relationship between ERMD and replicating nucleoids, we further examined the temporal relationship of active mtDNA synthesis to known mitochondrial division events: ER-linked mitochondrial constriction and recruitment of the mitochondrial division dynamin, DRP1 (*26*). We performed time-lapse microscopy to assess the relationship of POLG2-GFP–labeled nucleoids to mitochondrial division events in U2OS cells coexpressing mito-BFP and mCherry-DRP1. Consistent with our previous observations, a majority of mitochondrial division events marked by DRP1 assemblies occurred within 1 μm of a POLG2-GFP focus (Fig. 5, A and B). To further define the spatial link between replicating nucleoids and DRP1-marked division sites in live cells, we quantified the distance from the center of each POLG2-GFP focus to the position of matrix marker discontinuity, for 26 division events. We observed that division sites were enriched within a zone greater than 200 nm but less than 400 nm from a POLG2-GFP focus (fig. S4C). In every division event marked by mCherry-DRP1 (100%, $n$ = 26), POLG2-GFP–labeled nucleoids marked a future division site before mitochondrial constriction and also preceded the recruitment of mCherry-DRP1 to mitochondrial constrictions (Fig. 5B). Indeed, for some division events, POLG2-GFP was detected at a future division site more than 15 min before the recruitment of mCherry-DRP1. Thus, replication of mtDNA precedes known events linked to mitochondrial division.

### ER structure is required for mtDNA replication licensing and nucleoid distribution

The spatiotemporal relationship between replicating nucleoids and nascent ERMD sites prompted us to further examine the role of the ER in mtDNA replication and nucleoid distribution in cells. In addition to the nuclear envelope, the peripheral ER forms a dense and dynamic network of interconnected tubules and sheetlike structures (*37*–*39*). Members of the conserved reticulon protein family (RTN1, RTN2, RTN3, and RTN4/Nogo) are thought to be key factors in determining the ratio of ER tubules and sheetlike structures by localizing to and stabilizing highly curved ER tubules and edges of ER sheetlike structures (*40*, *41*). The coiled-coil membrane protein CLIMP63 is enriched in sheetlike ER structures where it is thought to maintain the lumenal distance between ER membrane bilayers (*41*–*43*). Overexpression of the ER-shaping proteins RTN4 or CLIMP63 in mammalian cells shifts the proportion of tubules to sheetlike structures or sheetlike structures to tubules, respectively (*40*, *41*, *44*). Thus, we used transient overexpression of CLIMP63 or RTN4A in COS-7 cells to manipulate ER structure and to examine its potential role in nucleoid replication and distribution. As expected, acute overexpression of fluorescently tagged RTN4A or CLIMP63 caused a dramatic increase in the proportion of tubular ER or sheetlike ER structures, respectively, as compared with control cells expressing Sec61b-mCherry, an ER membrane marker with no known role in membrane morphogenesis (Fig. 4A) (*40*). In contrast, and consistent with previous observations, ER morphology in cells simultaneously overexpressing CLIMP63-GFP and RTN4A-GFP was similar to control cells, consistent with a dynamic balance between the tubules and sheetlike structures within ER networks (Fig. 6A) (*41*).

In COS-7 labeled with MitoTracker Red and overexpressing CLIMP63, RTN4, or both, EdU pulse-labeling of mtDNA was used to assess the number of nucleoids engaged in mtDNA synthesis within a 4-hour window. After fixation, EdU-labeled nucleoids were detected in cells using AlexaFluor647, and total nucleoids and the ER were visualized using PicoGreen staining and indirect immunofluorescence with a Calreticulin antibody, respectively. Relative to control cells, there was a significant reduction in the number of EdU-labeled nucleoids in the cell population overexpressing CLIMP63-mCherry but not RTN4A-GFP (Fig. 6B). In addition, the fluorescence intensity




B

| ERMD events < 1um from POLG2-GFP | n = 15 |
|---|---|
| Observed | 73.3% |
| Expected | 22.1% |

**

**Fig. 3. Replicating nucleoids mark sites of ER-mediated division.** (**A**) Representative time-lapse images of a U2OS cell expressing mito-BFP, mRuby-KDEL (ER), and POLG2-GFP, demonstrating mitochondrial division at a mitochondrial-ER contact site spatially linked to POLG2-labeled nucleoid (arrowheads indicate division site). (**B**) Percentage of ERMD events (marked by mRuby-KDEL and mito-BFP as in Fig. 1B) in live U2OS cells that occurred within 1 μm of a POLG2-GFP–labeled nucleoid ($n$ = 15 events from seven cells, $^{**}P < 0.01$, Fisher's exact test). Significantly more division events occurred at preexisting POLG2-GFP foci (73.3%) than expected by random chance at stable ER-mitochondria contacts (22.1%). Scale bar, 2 μm.

of AlexaFluor647 at EdU-labeled nucleoids was significantly reduced in CLIMP63-overexpressing cells depleted of tubular ER in comparison with control and RTN4A-GFP–overexpressing cells (Fig. 6C). This observation indicates that the reduced number of apparent EdU foci in CLIMP63-overexpressing cells was not due to aggregation of replicating nucleoids. Moreover, in cells simultaneously overexpressing CLIMP63-mCherry and RTN4A-GFP, the number and fluorescence intensity of EdU-labeled nucleoids and ER morphology were similar to control cells, which suggested that the nucleoid phenotypes in CLIMP63-overexpressing cells were a consequence of changes in ER structure and not CLIMP63 expression per se (Fig. 6, B and C). Analysis of mtDNA copy number by qPCR indicated that there was no significant difference in mtDNA content in cells transiently overexpressing CLIMP63-mCherry, RTN4A-GFP, or both, relative to control cells (fig. S6A). This observation suggests that the transient reduction of the proportion of ER tubules in cells acutely disrupted mtDNA synthesis, as opposed to causing mtDNA loss or mtDNA EdU-labeling resistance (Fig. 6B). ER morphological phenotypes were not completely penetrant within the total population of cells overexpressing CLIMP63-GFP. We observed some cells with mosaic intracellular phenotypes in which the severity of ER tubule depletion varied spatially within the cytoplasm (Fig. 6, D and E). In this context, rare EdU-labeled nucleoids were detected even in cells highly overexpressing CLIMP63 and, in every case, were spatially linked to tubular ER colocalized with mitochondria (Fig. 6D, full field views in fig. S6B). Consistently, in live COS-7 cells overexpressing CLIMP63-mCherry and expressing POLG2-GFP and mito-BFP, POLG2-GFP–labeled nucleoids were observed adjacent to residual ER tubules colocalized with mitochondria (fig. S6C); these sites subsequently marked mitochondrial constriction and, ultimately, division events. In addition, we observed a reduction in the number of POLG2-GFP foci in cells overexpressing CLIMP63, as compared with control cells expressing the ER marker Sec61b-mCherry (fig. S6C). Thus, ER structure and, in particular, tubular ER-mitochondria contacts are necessary but not sufficient for homeostatic mtDNA replication.

Given the significant spatial link observed between ER-mitochondria contacts and nucleoid position (fig. S1A), we also examined the overall distribution of the total nucleoid population in cells under conditions of perturbed ER structure. In a majority of cells, overexpression of CLIMP63-mCherry, but not RTN4A-GFP, caused a significant increase in the area of resolvable TFAM-GFP foci, consistent with nucleoid aggregation and hence disturbed distribution (Fig. 6E and fig. S7, A and B). Further examination of a subset of CLIMP63-overexpressing cells that contained some normally distributed nucleoids revealed that these cells contained ER tubules colocalized with mitochondria at positions adjacent to TFAM-GFP–labeled nucleoids (fig. S7B). In addition, at the subcellular level, depletion of ER tubules was highly correlated with nucleoid aggregation (Fig. 6E). In cells simultaneously overexpressing CLIMP63-mCherry and RTN4A-GFP, the ER morphology and nucleoid distribution were similar to those in control cells (fig. S7A). Thus, normal ER structure is required for intracellular nucleoid distribution, which suggests that tubular ER-mitochondria contacts play a role in nucleoid segregation within the mitochondrial network of mammalian cells. These findings demonstrate critical functional interdependencies between mitochondrial and ER dynamics and mitochondrial genome maintenance.



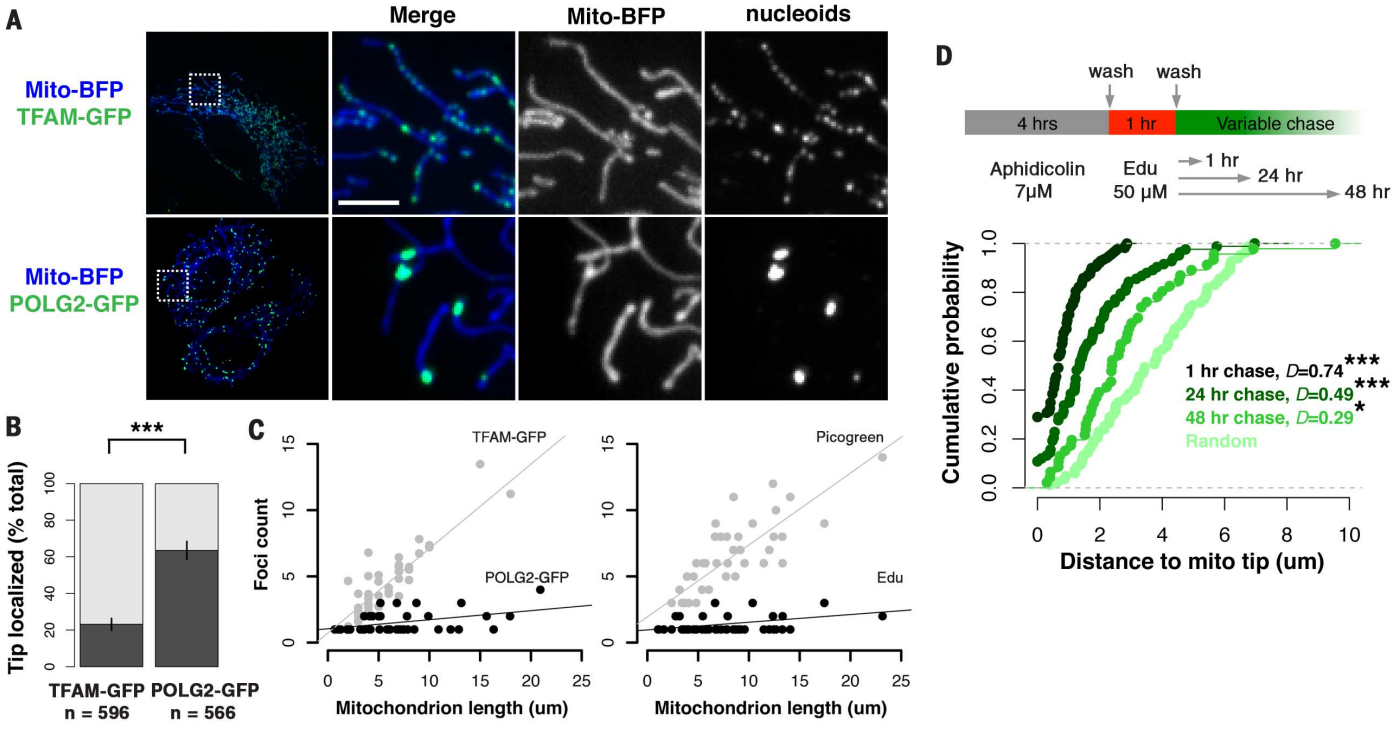



**Fig. 4. Nascent mtDNA is segregated to daughter mitochondria by division.** (**A**) Representative images of live U2OS cells expressing mito-BFP and TFAM-GFP (top) or POLG2-GFP (bottom). Scale bar, 5 μm. (**B**) Significant spatial enrichment of the total population of POLG2-GFP foci within 1 μm of mitochondrial tips, as compared with TFAM-GFP foci (in dark gray) in live U2OS cells as labeled and imaged in (A). (Data are means +/- SD, ***$P < 0.001$, two-tailed $t$ test.) (**C**) (Left) In live U2OS cells, the number of TFAM-GFP foci per mitochondrion, but not POLG2-GFP foci, is correlated with mitochondrion length. Linear regression with best-fit line is shown. For TFAM-GFP, adjusted $R^2 = 0.86$, Pearson's $R = 0.92$ ($P < 0.0001$). For POLG2-GFP, adjusted $R^2 = 0.15$, Pearson's $R = 0.41$ (n.s.). (Right) In fixed ARPE19 cells, the number of PicoGreen foci per mitochondrion, but not EdU foci, is correlated with mitochondrion length. Linear regression as on the left. For PicoGreen, adjusted $R^2 = 0.68$, Pearson's $R = 0.82$ ($P < 0.0001$). For EdU, adjusted $R^2 = 0.07$, Pearson's $R = 0.31$ (n.s.). (**D**) (Top) EdU pulse-chase experiments in ARPE19 cells. (Bottom) Empirical cumulative distribution analysis of EdU focus position along mitochondria, demonstrating depletion ($D$) of pulse-labeled nascent mtDNA from mitochondrial tips over time, toward a simulated random distribution (***$P < 0.001$, *$P < 0.05$, Kolmogorov-Smirnov test).

## Discussion

Our data indicate that within mitochondrial nucleoids in mammalian cells, homeostatic mtDNA synthesis is spatially linked to a small subset of ER-mitochondria contacts that are selectively coupled to mitochondrial division. We also observed that nucleoids at mitochondrial tips, which are the products of mitochondrial division, exhibit preferential motility within cells as compared with intramitochondrial nucleoids. We propose that the successive events of mtDNA replication, mitochondrial division, and mitochondrial motility are intimately linked and function together as part of a programmed process that ensures the accurate distribution of mtDNA within cells. Such a process may be especially important for highly polarized cells, such as neurons, whose long axonal processes likely depend on the transport and amplification of mitochondria and associated mtDNA derived from the cell body.

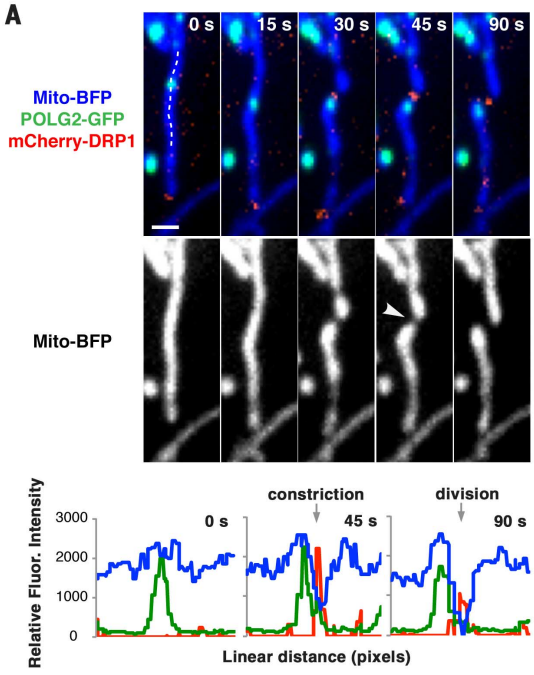
It will be important to determine the fundamental molecular mechanisms linking mtDNA replication initiation to mitochondrial division. Our analyses suggest that contacts between the ER and mitochondria are required to license mtDNA replication. Increasing evidence implicates interorganellar membrane contacts in the formation of membrane microdomains with specialized lipid and protein composition (*45*). In this capacity, ER-mitochondria contacts could function to facilitate the creation of a spatially defined platform within and on mitochondria that selectively recruits components required for the initiation of mtDNA replication, such as POLG1, POLG2, or other components of the mtDNA replisome. Supporting this idea are recent biochemical observations suggesting that mtDNA and mtDNA replisome proteins associate with cholesterol-rich membrane structures that would be predicted to have raft-like properties (*46*). Our findings also raise the question of how mtDNA replication, which occurs inside mitochondria, is coordinated with division events associated with the outer surface of the organelle. Perhaps, in addition to contact sites between mitochondria and the ER, there are intramitochondrial spatial determinants that contribute to division-site placement.

Our findings connect ER structure with mtDNA maintenance. This connection has implications for understanding the cellular pathology underlying human diseases and suggests that, for human diseases linked to defects in ER morphogenesis, pathogenesis could be a consequence of mitochondrial dysfunction (*47*, *48*).

## Methods

### Plasmids

All fluorescent protein constructs have been previously described. Mito-BFP and mCherry-DRP1 (*26*), mCherry-Sec61b (*49*), mCherry-CLIMP63 (*50*), GFP-CLIMP63, and RTN4A-GFP (*44*) were gifts from G. Voeltz. mRuby-KDEL was a gift from J. Wiedenmann (*51*). Human POLG2-GFP was a gift from W. Copeland (*34*). Human TFAM-GFP was a gift from M. Alexeyev (*52*).




B

| DRP1-marked divisions < 1um from POLG2-GFP | n = 26 |
|---|---|
| Observed | 76.9% |
| Expected | 25.6% |

**

**Fig. 5. Replicating nucleoids mark division sites before mitochondrial constriction or DRP1 recruitment.** (**A**) (Top) Time-lapse images of a U2OS cell expressing mito-BFP, mCherry-DRP1, and POLG2-GFP. Arrowhead indicates site of division. (Bottom) Line scan drawn along the mitochondrial tubule to show relative fluorescence intensity of mitochondria, DRP1 division machinery, and POLG2 signal for time points $t$ = 0 s (preconstriction), $t$ = 45 s (postconstriction), and $t$ = 90 s (postdivision). Scale bar, 1 μm. (**B**) The percentage of mitochondrial divisions [marked by mCherry-DRP1 as in (A)] that occur within 1 μm of a POLG2-GFP focus in live U2OS cells ($n$ = 26 events from 22 cells, $^{**}P < 0.01$, Fisher's exact test).

### Mammalian cell growth and transfection

U2OS, COS-7, and ARPE19 cells (ATCC) were grown in high-glucose Dulbecco's Modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (FBS) and 1% penicillin/streptomycin. Cells were seeded at ~$0.5 \times 10^5$ cells per ml in 35-mm glass-bottom dishes (MatTek) 24 hours before transient transfection and 40 hours before imaging. Plasmid transfections were performed for 4 hours in serum- and antibiotic-free DMEM with 3 μl FuGENE6 reagent (Millipore) per dish. Sixteen hours later, cells were imaged in Fluorobrite DMEM (ThermoFisher) containing 10% FBS.

### Cell fixation, antibodies, and immunofluorescence

Cells were seeded as described above; 24 hours later, cells were stained with 500 nM MitoTracker Red chloromethyl-X-rosamine (CMX-Ros) (ThermoFisher), and where indicated, a 1:1000 dilution of PicoGreen DNA stain (ThermoFisher) for 15 min at 37°C. Cells were rinsed once in complete medium and twice in warm phosphate-buffered saline (PBS) and were fixed in 4% paraformaldehyde in PBS pH 7.4 for 20 min at room temperature. Dishes were washed twice in PBS and permeabilized in 0.1% TritonX-100 for 20 min. Dishes were blocked in 3% bovine serum albumin (BSA) PBS solution for 1 hour at room temperature. Primary antibodies were added at 1:1000 dilution in PBST (PBS pH 7.4, 1% BSA, 0.1% Tween-20) overnight at 4°C, rinsed twice in PBS, and incubated with Alexa-Fluor–conjugated secondary antibodies at 1:2000 dilution in PBST for 1 hour. Antibodies used: mouse anti-GFP AlexaFluor 488 conjugate (ThermoFisher), donkey anti-rabbit AlexaFluor 405 conjugate (ThermoFisher), anti-GFP (clone N86/8, Neuromabs), rabbit anti-Calreticulin (2907, Abcam).

EdU incorporation was detected via Click-iT EdU AlexaFluor 647 labeling kit (C10640, ThermoFisher) according to the manufacturer's instructions with minor deviations. Briefly, cells were incubated in 7 μM aphidicolin (A4487, Sigma) for 4 hours in complete medium before a pulse of 50 μM EdU. The EdU pulse was followed by a chase in EdU-free complete DMEM for 1, 24, or 48 hours as described in the main text. For all EdU-labeling experiments, cells were fixed and stained while subconfluent, during logarithmic growth.

### Spinning-disk confocal microscopy

Live cell imaging was performed using the spinning-disk module of an inverted objective fluorescence microscope [Marianas spinning-disk confocal (SDC) real-time 3D Confocal-TIRF (total internal reflection) microscope; Intelligent Imaging Innovations] with 100×, 1.46 numerical aperture objective. Either a Photometrics QuantiEM electron multiplying charge-coupled device or Hamamatsu Orca Flash 4.0 scientific complementary metal–oxide–semiconductor (sCMOS) camera was used, depending on the experiment. Images were captured with Slidebook (Intelligent Imaging Innovations); if necessary, linear adjustments were made with ImageJ (NIH). Morphological and quantification analyses were performed in Nikon Elements-Advanced Research (Nikon) as described below. Scale bars were generated using Slidebook.

### Mitochondrial division proximity analysis

To predict the frequency that mitochondrial division could occur at POLG2-GFP–marked ER-mitochondrial contacts by random chance, we counted the total number of persistent contacts from each mitochondrion containing a POLG2-GFP focus in the frames leading up to each division

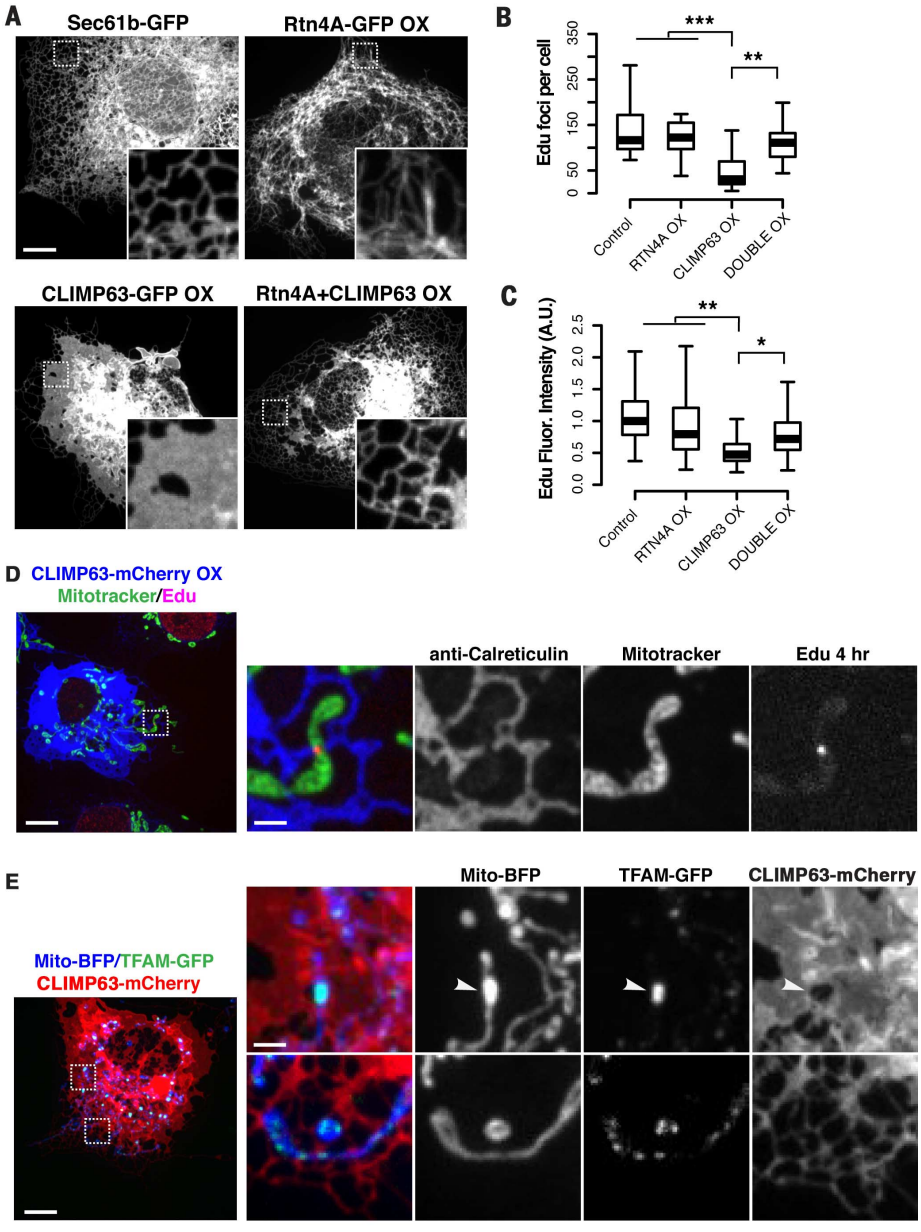



**Fig. 6. ER tubules license mtDNA synthesis and are required for nucleoid distribution.** (**A**) ER network morphologies in representative COS-7 cells expressing fluorescently tagged ER membrane proteins: Sec61b-GFP (top left), RTN4A-GFP (top right), CLIMP63-GFP overexpression (OX) (bottom left), and RTN4A-GFP and CLIMP63-mCherry double overexpression (bottom right). Inset regions magnified 5×. (**B**) Quantification of the number of mitochondrial EdU foci per fixed COS-7 cell after a 4-hour pulse of 50 μM EdU, in cells labeled with MitoTracker Red and indicated ER markers, $n = 15+$ cells per condition (***$P < 0.001$, **$P < 0.01$, two-tailed $t$ test). (**C**) Quantification of fluorescence intensity of mitochondrial EdU foci, $n = 600+$ foci from 15+ cells per condition (**$P < 0.01$; *$P < 0.05$, two-tailed $t$ test). (**D**) Image of fixed COS-7 cell overexpressing CLIMP63-mCherry (ER) after a 4-hour pulse of 50 μM EdU in cells labeled with MitoTracker Red. (**E**) Image of a live COS-7 cell expressing mito-BFP (mitochondria), TFAM-GFP (nucleoids), and overexpressing CLIMP63-mCherry (ER). (Left) Full field view of entire cell; (right) examples of aggregated nucleoids associated with sheetlike ER (top) and distributed nucleoids associated with reticular ER (bottom). Arrowheads indicate colocalization. Scale bars: (A), (D), and (E), 10 μm; insets in (D) and (E), 2 μm.

event, considering each to be a potential future division site. We then compared the number of ERMD events that occurred within 1 μm of the POLG2-GFP to the total number of persistent contacts and averaged over all events. For example, a mitochondrion with three persistent contacts and one division event would have a per-contact expected frequency of one-third.

### MtDNA copy number and qPCR analyses

Total DNA was isolated from U2OS cells by using the DNeasy Blood and Tissue Kit (Qiagen). Quantitative PCR was carried out using SsoAdvanced universal probes supermix (Biorad). Mitochondrial DNA copy number of control, and TFAM-, CLIMP63-, RTN4A- and double-overexpression osteosarcoma cells was performed as in (*11*) by using primer sequences described therein. Briefly, mtDNA copy number was normalized to nuclear DNA by amplifying an ~132-nucleotide fragment of cytochrome b as compared with a similarly sized fragment of the single-copy nuclear gene, APP. The delta delta Ct method (the ratio of our target gene in our treated sample relative to our untreated sample change in measured cycle thresholds) was used in calculations of fold change.

### Statistical analyses, plotting, and modeling

All statistical analyses and plotting were performed in R version 3.2.0 within the RStudio development environment, version 0.99.441. To model the distribution of EdU foci in fixed ARPE19 cells, we used the empirical cumulative distribution function ecdf(). Random simulated data were generated with the runif() function by using 83 observations and a range of 0 to 7, consistent with our real data from the first biological replicate of the 1-hour EdU pulse with 1-hour chase. To recreate the random data set shown in Fig. 4D, use the set.seed() function as follows:

```
>set.seed(100)#random number generator, initial state
>randomDistToTip<-runif(83, min = 0, max = 7)#generate random dataset
>R<-ecdf(randomDistToTip)#calculate ecdf
>plot(R)#plot random dataset
```

### REFERENCES AND NOTES


1. W. C. Copeland, Defects of mitochondrial DNA replication. *J. Child Neurol.* **29**, 1216–1224 (2014). doi: 10.1177/0883073814537380; pmid: 24985751
2. J. Nunnari, A. Suomalainen, Mitochondria: In sickness and in health. *Cell* **148**, 1145–1159 (2012). doi: 10.1016/j.cell.2012.02.035; pmid: 22424226
3. A. Trifunovic *et al.*, Premature ageing in mice expressing defective mitochondrial DNA polymerase. *Nature* **429**, 417–423 (2004). doi: 10.1038/nature02517; pmid: 15164064
4. G. C. Kujoth *et al.*, Mitochondrial DNA mutations, oxidative stress, and apoptosis in mammalian aging. *Science* **309**, 481–484 (2005). doi: 10.1126/science.1112125; pmid: 16020738
5. N. Garrido *et al.*, Composition and dynamics of human mitochondrial nucleoids. *Mol. Biol. Cell* **14**, 1583–1596 (2003). doi: 10.1091/mbc.E02-07-0399; pmid: 12686611
6. P. A. Ropp, W. C. Copeland, Cloning and characterization of the human mitochondrial DNA polymerase, DNA polymerase gamma. *Genomics* **36**, 449–458 (1996). doi: 10.1006/geno.1996.0490; pmid: 8884268
7. S. E. Lim, M. J. Longley, W. C. Copeland, The mitochondrial p55 accessory subunit of human DNA polymerase gamma enhances DNA binding, promotes processive DNA synthesis, and confers *N*-ethylmaleimide resistance. *J. Biol. Chem.* **274**, 38197–38203 (1999). doi: 10.1074/jbc.274.53.38197; pmid: 10608893
8. L. S. Kaguni, DNA polymerase γ, the mitochondrial replicase. *Annu. Rev. Biochem.* **73**, 293–320 (2004). doi: 10.1146/annurev.biochem.72.121801.161455; pmid: 15189144
9. S. Wanrooij, S. Goffart, J. L. O. Pohjoismäki, T. Yasukawa, J. N. Spelbrink, Expression of catalytic mutants of the mtDNA helicase Twinkle and polymerase POLG causes distinct replication stalling phenotypes. *Nucleic Acids Res.* **35**, 3238–3251 (2007). doi: 10.1093/nar/gkm215; pmid: 17452351
10. S. Wanrooij, M. Falkenberg, The human mitochondrial replication fork in health and disease. *Biochim. Biophys. Acta*

**1797**, 1378–1388 (2010). doi: 10.1016/j.bbabio.2010.04.015; pmid: 20417176
11. H. Tyynismaa *et al*., Twinkle helicase is essential for mtDNA maintenance and regulates mtDNA copy number. *Hum. Mol. Genet.* **13**, 3219–3227 (2004). doi: 10.1093/hmg/ddh342; pmid: 15509589
12. J. A. Korhonen, X. H. Pham, M. Pellegrini, M. Falkenberg, Reconstitution of a minimal mtDNA replisome in vitro. *EMBO J.* **23**, 2423–2429 (2004). doi: 10.1038/sj.emboj.7600257; pmid: 15167897
13. T. I. Alam *et al*., Human mitochondrial DNA is packaged with TFAM. *Nucleic Acids Res.* **31**, 1640–1645 (2003). doi: 10.1093/nar/gkg251; pmid: 12626705
14. B. A. Kaufman *et al*., The mitochondrial transcription factor TFAM coordinates the assembly of multiple DNA molecules into nucleoid-like structures. *Mol. Biol. Cell* **18**, 3225–3236 (2007). doi: 10.1091/mbc.E07-05-0404; pmid: 17581862
15. H. B. Ngo, J. T. Kaiser, D. C. Chan, The mitochondrial transcription and packaging factor Tfam imposes a U-turn on mitochondrial DNA. *Nat. Struct. Mol. Biol.* **18**, 1290–1296 (2011). doi: 10.1038/nsmb.2159; pmid: 22037171
16. C. Kukat *et al*., Super-resolution microscopy reveals that mammalian mitochondrial nucleoids have a uniform size and frequently contain a single copy of mtDNA. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 13534–13539 (2011). doi: 10.1073/pnas.1109263108; pmid: 21808029
17. C. Kukat *et al*., Cross-strand binding of TFAM to a single mtDNA molecule forms the mitochondrial nucleoid. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 11288–11293 (2015). doi: 10.1073/pnas.1512131112; pmid: 26305956
18. T. A. Brown *et al*., Superresolution fluorescence imaging of mitochondrial nucleoids reveals their spatial range, limits, and membrane interaction. *Mol. Cell. Biol.* **31**, 4994–5010 (2011). doi: 10.1128/MCB.05694-11; pmid: 22006021
19. C. W. Birky Jr., The inheritance of genes in mitochondria and chloroplasts: Laws, mechanisms, and models. *Annu. Rev. Genet.* **35**, 125–148 (2001). doi: 10.1146/annurev.genet.35.102401.090231; pmid: 11700280
20. J. L. O. Pohjoismäki *et al*., Human heart mitochondrial DNA is organized in complex catenated networks containing abundant four-way junctions and replication forks. *J. Biol. Chem.* **284**, 21446–21457 (2009). doi: 10.1074/jbc.M109.016600; pmid: 19525233
21. J. Magnusson, M. Orth, P. Lestienne, J.-W. Taanman, Replication of mitochondrial DNA occurs throughout the mitochondria of cultured human cells. *Exp. Cell Res.* **289**, 133–142 (2003). doi: 10.1016/S0014-4827(03)00249-0; pmid: 12941611
22. J. Nunnari *et al*., Mitochondrial transmission during mating in *Saccharomyces cerevisiae* is determined by mitochondrial fusion and fission and the intramitochondrial segregation of mitochondrial DNA. *Mol. Biol. Cell* **8**, 1233–1242 (1997). doi: 10.1091/mbc.8.7.1233; pmid: 9243504
23. K. Labbé, A. Murley, J. Nunnari, Determinants and functions of mitochondrial behavior. *Annu. Rev. Cell Dev. Biol.* **30**, 357–391 (2014). doi: 10.1146/annurev-cellbio-101011-155756; pmid: 25288115
24. E. Smirnova, L. Griparic, D. L. Shurland, A. M. van der Bliek, Dynamin-related protein Drp1 is required for mitochondrial division in mammalian cells. *Mol. Biol. Cell* **12**, 2245–2256 (2001). doi: 10.1091/mbc.12.8.2245; pmid: 11514614
25. E. Smirnova, D. L. Shurland, S. N. Ryazantsev, A. M. van der Bliek, A human dynamin-related protein controls the distribution of mitochondria. *J. Cell Biol.* **143**, 351–358 (1998). doi: 10.1083/jcb.143.2.351; pmid: 9786947
26. J. R. Friedman *et al*., ER tubules mark sites of mitochondrial division. *Science* **334**, 358–362 (2011). doi: 10.1126/science.1207385; pmid: 21885730
27. R. Ban-Ishihara, T. Ishihara, N. Sasaki, K. Mihara, N. Ishihara, Dynamics of nucleoid structure regulated by mitochondrial fission contributes to cristae reformation and release of cytochrome c. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 11863–11868 (2013). doi: 10.1073/pnas.1301951110; pmid: 23821750
28. T. Ishihara *et al*., Dynamics of mitochondrial DNA nucleoids regulated by mitochondrial fission is essential for maintenance of homogeneously active mitochondria during neonatal heart development. *Mol. Cell. Biol.* **35**, 211–223 (2015). doi: 10.1128/MCB.01054-14; pmid: 25348719
29. K. Itoh, Y. Tamura, M. Iijima, H. Sesaki, Effects of Fcj1-Mos1 and mitochondrial division on aggregation of mitochondrial DNA nucleoids and organelle morphology. *Mol. Biol. Cell* **24**, 1842–1851 (2013). doi: 10.1091/mbc.E13-03-0125; pmid: 23615445
30. A. Murley *et al*., ER-associated mitochondrial division links the distribution of mitochondria and mitochondrial DNA in yeast. *eLife* **2**, e00422 (2013). pmid: 23682313
31. F. J. Iborra, H. Kimura, P. R. Cook, The functional organization of mitochondrial genomes in human cells. *BMC Biol.* **2**, 9 (2004). doi: 10.1186/1741-7007-2-9; pmid: 15157274
32. N. Rajala, J. M. Gerhold, P. Martinsson, A. Klymov, J. N. Spelbrink, Replication factors transiently associate with mtDNA at the mitochondrial inner membrane to facilitate replication. *Nucleic Acids Res.* **42**, 952–967 (2014). doi: 10.1093/nar/gkt988; pmid: 24163258
33. S. Meeusen, J. Nunnari, Evidence for a two membrane-spanning autonomous mitochondrial DNA replisome. *J. Cell Biol.* **163**, 503–510 (2003). doi: 10.1083/jcb.200304040; pmid: 14597773
34. M. J. Young, M. M. Humble, K. L. DeBalsi, K. Y. Sun, W. C. Copeland, POLG2 disease variants: Analyses reveal a dominant negative heterodimer, altered mitochondrial localization and impaired respiratory capacity. *Hum. Mol. Genet.* **24**, 5184–5197 (2015). doi: 10.1093/hmg/ddv240; pmid: 26123486
35. A. Salic, T. J. Mitchison, A chemical method for fast and sensitive detection of DNA synthesis in vivo. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 2415–2420 (2008). doi: 10.1073/pnas.0712168105; pmid: 18272492
36. T. Kanki *et al*., Architectural role of mitochondrial transcription factor A in maintenance of human mitochondrial DNA. *Mol. Cell. Biol.* **24**, 9823–9834 (2004). doi: 10.1128/MCB.24.22.9823-9834.2004; pmid: 15509786
37. J. Hu, W. A. Prinz, T. A. Rapoport, Weaving the web of ER tubules. *Cell* **147**, 1226–1231 (2011). doi: 10.1016/j.cell.2011.11.022; pmid: 22153070
38. G. K. Voeltz, M. M. Rolls, T. A. Rapoport, Structural organization of the endoplasmic reticulum. *EMBO Rep.* **3**, 944–950 (2002). doi: 10.1093/embo-reports/kvf202; pmid: 12370207
39. M. West, N. Zurek, A. Hoenger, G. K. Voeltz, A 3D analysis of yeast ER structure reveals how ER domains are organized by membrane curvature. *J. Cell Biol.* **193**, 333–346 (2011). doi: 10.1083/jcb.201011039; pmid: 21502358
40. G. K. Voeltz, W. A. Prinz, Y. Shibata, J. M. Rist, T. A. Rapoport, A class of membrane proteins shaping the tubular endoplasmic reticulum. *Cell* **124**, 573–586 (2006). doi: 10.1016/j.cell.2005.11.047; pmid: 16469703
41. Y. Shibata *et al*., Mechanisms determining the morphology of the peripheral ER. *Cell* **143**, 774–788 (2010). doi: 10.1016/j.cell.2010.11.007; pmid: 21111237
42. D. R. Klopfenstein *et al*., Subdomain-specific localization of CLIMP-63 (p63) in the endoplasmic reticulum is mediated by its luminal alpha-helical segment. *J. Cell Biol.* **153**, 1287–1300 (2001). pmid: 11402071
43. T. Cui-Wang *et al*., Local zones of endoplasmic reticulum complexity confine cargo in neuronal dendrites. *Cell* **148**, 309–321 (2012). doi: 10.1016/j.cell.2011.11.056; pmid: 22265418
44. Y. Shibata *et al*., The reticulon and DP1/Yop1p proteins form immobile oligomers in the tubular endoplasmic reticulum. *J. Biol. Chem.* **283**, 18892–18904 (2008). doi: 10.1074/jbc.M800986200; pmid: 18442980
45. A. Murley, J. Nunnari, The emerging network of mitochondria-organelle contacts. *Mol. Cell* **61**, 648–653 (2016). doi: 10.1016/j.molcel.2016.01.031; pmid: 26942669
46. J. M. Gerhold *et al*., Human mitochondrial DNA-protein complexes attach to a cholesterol-rich membrane structure. *Sci. Rep.* **5**, 15292 (2015). doi: 10.1038/srep15292; pmid: 26478270
47. C. Blackstone, Cellular pathways of hereditary spastic paraplegia. *Annu. Rev. Neurosci.* **35**, 25–47 (2012). doi: 10.1146/annurev-neuro-062111-150400; pmid: 22540978
48. E. I. Rugarli, T. Langer, Translating m-AAA protease function in mitochondria to hereditary spastic paraplegia. *Trends Mol. Med.* **12**, 262–269 (2006). doi: 10.1016/j.molmed.2006.04.002; pmid: 16647881

## ACKNOWLEDGMENTS

We thank members of the Nunnari lab and W. Copeland for discussion, and M. Paddy and the University of California (UC), Davis, Department of Molecular and Cellular Biology Microscopy Facility for helpful suggestions. This work is supported by NIH grants R01GM106019 and R01GM097432 (to J.N.), an NIH Ruth L. Kirschstein Postdoctoral Fellowship F32GM113388 (to S.C.L.), and a UC Davis Provost's Undergraduate Research Fellowship (to L.F.U.). J.N. is on the Advisory Board of Mitobridge and declares no financial interest related to this work. The other authors declare that no competing interests exist. All data are included in the main manuscript and supplementary materials.

## SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/353/6296/aaf5549/suppl/DC1
Figs. S1 to S7
References (*49*–*52*)

26 February 2016; accepted 26 May 2016
10.1126/science.aaf5549

RESEARCH ARTICLE SUMMARY

VOLCANOLOGY

# Gradual caldera collapse at Bárdarbunga volcano, Iceland, regulated by lateral magma outflow

**Magnús T. Gudmundsson,* Kristín Jónsdóttir, Andrew Hooper, Eoghan P. Holohan, Sæmundur Ari Halldórsson, Benedikt G. Ófeigsson, Simone Cesca, Kristín S. Vogfjörd, Freysteinn Sigmundsson, Thórdís Högnadóttir, Páll Einarsson, Olgeir Sigmarsson, Alexander H. Jarosch, Kristján Jónasson, Eyjólfur Magnússon, Sigrún Hreinsdóttir, Marco Bagnardi, Michelle M. Parks, Vala Hjörleifsdóttir, Finnur Pálsson, Thomas R. Walter, Martin P. J. Schöpfer, Sebastian Heimann, Hannah I. Reynolds, Stéphanie Dumont, Eniko Bali, Gudmundur H. Gudfinnsson, Torsten Dahm, Matthew J. Roberts, Martin Hensch, Joaquín M. C. Belart, Karsten Spaans, Sigurdur Jakobsson, Gunnar B. Gudmundsson, Hildur M. Fridriksdóttir, Vincent Drouin, Tobias Dürig, Guðfinna Aðalgeirsdóttir, Morten S. Riishuus, Gro B. M. Pedersen, Tayo van Boeckel, Björn Oddsson, Melissa A. Pfeffer, Sara Barsotti, Baldur Bergsson, Amy Donovan, Mike R. Burton, Alessandro Aiuppa**

**INTRODUCTION:** The Bárdarbunga caldera volcano in central Iceland collapsed from August 2014 to February 2015 during the largest eruption in Europe since 1784. An ice-filled subsidence bowl, 110 square kilometers ($km^2$) in area and up to 65 meters (m) deep developed, while magma drained laterally for 48 km along a subterranean path and erupted as a major lava flow northeast of the volcano. Our data provide unprecedented insight into the workings of a collapsing caldera.

**RATIONALE:** Collapses of caldera volcanoes are, fortunately, not very frequent, because they are often associated with very large volcanic eruptions. On the other hand, the rarity of caldera collapses limits insight into this major geological hazard. Since the formation of Katmai caldera in 1912, during the 20th century's largest eruption, only five caldera collapses are known to have occurred before that at Bárdarbunga. We used aircraft-based altimetry, satellite photogrammetry, radar interferometry, ground-based GPS, evolution of seismicity, radio-echo soundings of ice thickness, ice flow modeling, and geobarometry to describe and analyze the evolving subsidence geometry, its underlying cause, the amount of magma erupted, the geometry of the subsurface caldera ring faults, and the moment tensor solutions of the collapse-related earthquakes.

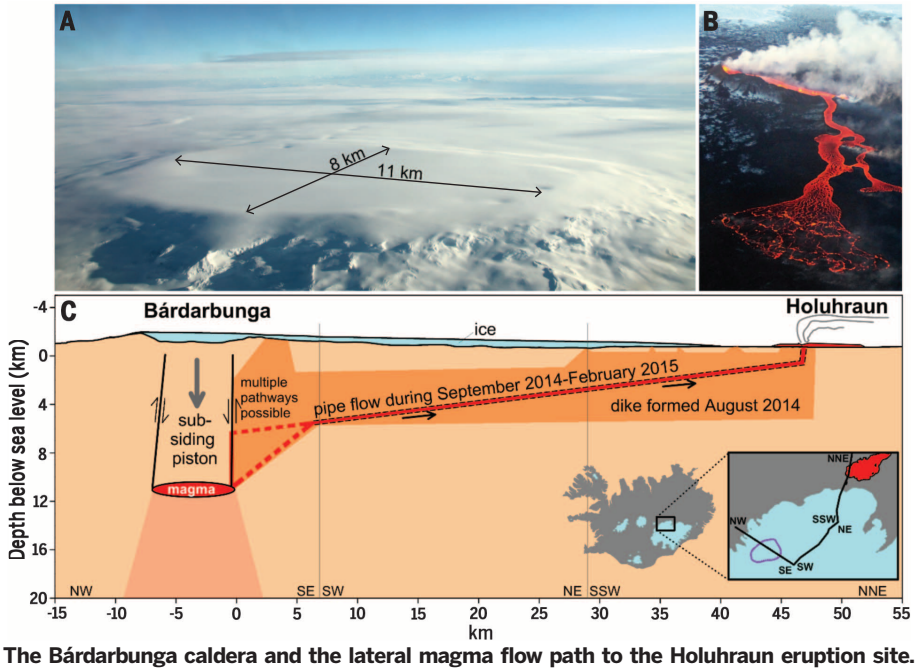
**RESULTS :** After initial lateral withdrawal of magma for some days though a magma-filled fracture propagating through Earth's upper crust, preexisting ring faults under the volcano were reactivated over the period 20 to 24 August, marking the onset of collapse. On 31 August, the eruption started, and it terminated when the collapse stopped, having produced 1.5 km of basaltic lava. The subsidence of the caldera declined with time in a near-exponential manner, in phase with the lava flow rate.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aaf8988

The volume of the subsidence bowl was about 1.8 $km^3$. Using radio-echo soundings, we find that the subglacial bedrock surface after the collapse is down-sagged, with no indications of steep fault escarpments. Using geobarometry, we determined the depth of the magma reservoir to be ~12 km, and modeling of geodetic observations gives a similar result. High-precision earthquake locations and moment tensor analysis of the remarkable magnitude *M*5 earthquake series are consistent with steeply dipping ring faults. Statistical analysis of seismicity reveals communication over tens of kilometers between the caldera and the dike.

**CONCLUSION:** We conclude that interaction between the pressure exerted by the subsiding reservoir roof and the physical properties of the subsurface flow path explain the gradual near-exponential decline of both the collapse rate and the intensity of the 180-day-long eruption. By combining our various data sets, we show that the onset of collapse was caused by outflow of magma from underneath the caldera when 12 to 20% of the total magma intruded and erupted had flowed from the magma reservoir. However, the continued subsidence was driven by a feedback between the pressure of the piston-like block overlying the reservoir and the 48-km-long magma outflow path. Our data provide better constraints on caldera mechanisms than previously available, demonstrating what caused the onset and how both the roof overburden and the flow path properties regulate the collapse. ■




**The Bárdarbunga caldera and the lateral magma flow path to the Holuhraun eruption site.** (**A**) Aerial view of the ice-filled Bárdarbunga caldera on 24 October 2014, view from the north. (**B**) The effusive eruption in Holuhraun, about 40 km to the northeast of the caldera. (**C**) A schematic cross section through the caldera and along the lateral subterranean flow path between the magma reservoir and the surface.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: mtg@hi.is
Cite this article as M. T. Gudmundsson *et al.*, *Science* **353**, aaf8988 (2016). DOI: 10.1126/science.aaf8988.

RESEARCH ARTICLE

VOLCANOLOGY

# Gradual caldera collapse at Bárdarbunga volcano, Iceland, regulated by lateral magma outflow

Magnús T. Gudmundsson,[1]* Kristín Jónsdóttir,[2] Andrew Hooper,[3] Eoghan P. Holohan,[4,5] Sæmundur A. Halldórsson,[1] Benedikt G. Ófeigsson,[2] Simone Cesca,[4] Kristín S. Vogfjörd,[2] Freysteinn Sigmundsson,[1] Thórdís Högnadóttir,[1] Páll Einarsson,[1] Olgeir Sigmarsson,[1,6] Alexander H. Jarosch,[1] Kristján Jónasson,[7] Eyjólfur Magnússon,[1] Sigrún Hreinsdóttir,[8] Marco Bagnardi,[3] Michelle M. Parks,[1] Vala Hjörleifsdóttir,[9] Finnur Pálsson,[1] Thomas R. Walter,[4] Martin P. J. Schöpfer,[10] Sebastian Heimann,[4] Hannah I. Reynolds,[1] Stéphanie Dumont,[1] Eniko Bali,[1] Gudmundur H. Gudfinnsson,[1] Torsten Dahm,[4] Matthew J. Roberts,[2] Martin Hensch,[2] Joaquín M. C. Belart,[1] Karsten Spaans,[3] Sigurdur Jakobsson,[1] Gunnar B. Gudmundsson,[2] Hildur M. Fridriksdóttir,[1,2] Vincent Drouin,[1] Tobias Dürig,[1] Guðfinna Aðalgeirsdóttir,[1] Morten S. Riishuus,[1] Gro B. M. Pedersen,[1] Tayo van Boeckel,[1] Björn Oddsson,[11] Melissa A. Pfeffer,[2] Sara Barsotti,[2] Baldur Bergsson,[2] Amy Donovan,[12] Mike R. Burton,[13] Alessandro Aiuppa[14]

Large volcanic eruptions on Earth commonly occur with a collapse of the roof of a crustal magma reservoir, forming a caldera. Only a few such collapses occur per century, and the lack of detailed observations has obscured insight into the mechanical interplay between collapse and eruption. We use multiparameter geophysical and geochemical data to show that the 110-square-kilometer and 65-meter-deep collapse of Bárdarbunga caldera in 2014–2015 was initiated through withdrawal of magma, and lateral migration through a 48-kilometers-long dike, from a 12-kilometers deep reservoir. Interaction between the pressure exerted by the subsiding reservoir roof and the physical properties of the subsurface flow path explain the gradual, near-exponential decline of both collapse rate and the intensity of the 180-day- long eruption.

Calderas are 1- to 100-km-diameter depressions found in volcanic regions of Earth and other planets. They mainly form by collapse of overburden into a subterranean magma reservoir during large volcanic eruptions, including the largest known supereruptions (*1*–*8*). From 1900 to 2014 CE, only six cases have been documented, with varying degrees of detail. The collapses of Katmai in 1912 and Pinatubo in 1991 occurred during explosive silicic (andesite-rhyolite) eruptions, the largest of the 20th century. The collapses of Fernandina in 1968, Tolbachik in 1975–1976, Miyakejima in 2000, and Piton de la Fournaise in 2007 were associated with mainly effusive mafic (basalt–basaltic andesite) intrusive activity and eruptions (*2*, *9*–*12*).

The consensus from field and modeling studies is that caldera collapse progresses from initial surface down-sag to fault-controlled subsidence (*1*, *8*, *13*, *14*). The limited number of modern examples and the scarcity of geophysical data leaves open the question of whether collapse occurs suddenly or gradually during the course of an eruption. The issue of whether collapse drives magma movement and eruption or eruption drives collapse also remains unresolved. Previous geological, geophysical, and modeling studies have produced a diverse and inconsistent set of answers to such questions (*2*, *4*, *15*, *16*). The caldera collapse at Bárdarbunga in central Iceland from August 2014 to February 2015 offers a unique opportunity to address them directly.

## The Bárdarbunga volcano and the Holuhraun eruption of 2014–2015

Bárdarbunga volcano (Fig. 1) and its related fissure swarms form a 150-km-long volcanic system on the boundary between the North American and Eurasian tectonic plates. The volcano resides beneath the Vatnajökull ice cap and has a broadly elliptic 11- by 8-km-wide and 500- to 700-m-deep caldera with a long axis trending east-northeast. About 700 to 800 m of ice fills the caldera (*17*, *18*). More than 20 eruptions have occurred on the fissure swarms outside the caldera in the last 12 centuries, including three that produced 1 to 4 $km^3$ of magma, but no eruptions are known within the caldera in this period (*19*).

At 4 UTC on 16 August 2014, the onset of intense seismicity beneath the caldera marked the beginning of a major rifting event (*20*). The seismic activity was mostly located in the southeast corner of the caldera in the first few hours, but it soon began to propagate out of the caldera toward the southeast (Fig. 2). After propagating to about 7 km from the caldera rim, 15 hours after the onset of seismicity (~19 UTC), the moving earthquake cluster took a 90° turn and started migrating toward the northeast. In the 2 weeks that followed, surface deformation and migration of seismicity indicated that a magmatic dike propagated laterally northeastward for a futher 41 km in the uppermost 6 to 10 km of Earth's crust (*20*, *21*). On 31 August, a major effusive eruption began above the far end of the dike; this lasted 6 months and produced 1.5 ± 0.2 $km^3$ of lava (~1.4 ± 0.2 $km^3$ of bubble-free magma) (*22*), making it the largest in Iceland (or Europe) since the 1783–1784 Laki eruption. Combined with the 0.5 ± 0.1 $km^3$ dike (*20*), the total volume of identified intruded and erupted magma was 1.9 ± 0.3 $km^3$.

## The onset of collapse

After the initial seismic activity in the caldera receded late on 16 August, seismicity was relatively minor there until 20 August. At the same time, our Global Positioning System (GPS) time series from stations close to the caldera suggest that deflation of the magma reservoir started on 16 August (*20*). On 20 August, caldera seismicity increased progressively, with a series of earthquakes of magnitude *M*4 to *M*5.8 occurring in the following days (Fig. 2). The first two events occurred on the southern caldera rim (*M*4.7 on 20 August and *M*5.1 on 21 August). After these earthquakes, three events of similar magnitude occurred on the northern rim on 23 August, followed by four events on the southern rim on 24 to 25 August. On 26 August, activity shifted again to the northern rim with a *M*5.8 earthquake, the largest in the whole series. These data indicate that substantial movement on ring faults started on the south side with the 20 to 21 August earthquakes, then began on the north side on 23 August, and by 24 August the ring faults on both sides where slipping, a process that did not terminate until at the end of February. Onset of collapse therefore likely occurred on 20 August, with the ring fault fully activated on 24 August. If we compare the evolution of the dike together with the seismic moment release of the caldera collapse earthquakes, we can clearly see that the dike migration leads the moment release curve (Fig. 2A). We therefore conclude that onset of

[1]Nordvulk, Institute of Earth Sciences, University of Iceland, Sturlugata 7, 101 Reykjavík, Iceland. [2]Icelandic Meteorological Office, IS-150 Reykjavík, Iceland. [3]Centre for the Observation and Modelling of Earthquakes, Volcanoes and Tectonics (COMET), School of Earth and Environment, University of Leeds, Leeds LS2 9JT, UK. [4]GFZ German Research Centre for Geosciences, Telegrafenberg, 14473 Potsdam, Germany. [5]UCD School of Earth Sciences, University College Dublin, Ireland. [6]Laboratoire Magmas et Volcans, CNRS-Université Blaise Pascal-IRD, 63038 Clermont-Ferrand, France. [7]Faculty of Industrial and Mechanical Engineering and Computer Science, University of Iceland, Hjarðarhagi 2-6, 107 Reykjavík, Iceland. [8]GNS Science, Post Office Box 30368, Lower Hutt 5040, New Zealand. [9]Departamento de Sismología, Instituto de Geofísica, Universidad Nacional Autónoma de Mexico, 04510 Ciudad de México, Mexico. [10]Department for Geodynamics and Sedimentology, University of Vienna, Althanstrasse 14, A-1090 Vienna, Austria. [11]National Commissioner of the Icelandic Police, Department of Civil Protection and Emergency Management, Skúlagata 21, 101 Reykjavík, Iceland. [12]King's College London, King's Building, Strand Campus, London WC2R 2LS, England, UK. [13]University of Manchester, Williamson Building, Oxford Road, Manchester, M13 9PL, UK. [14]University of Palermo, Piazza Marina, 61 90133, Palermo, Italy.
*Corresponding author. Email: mtg@hi.is

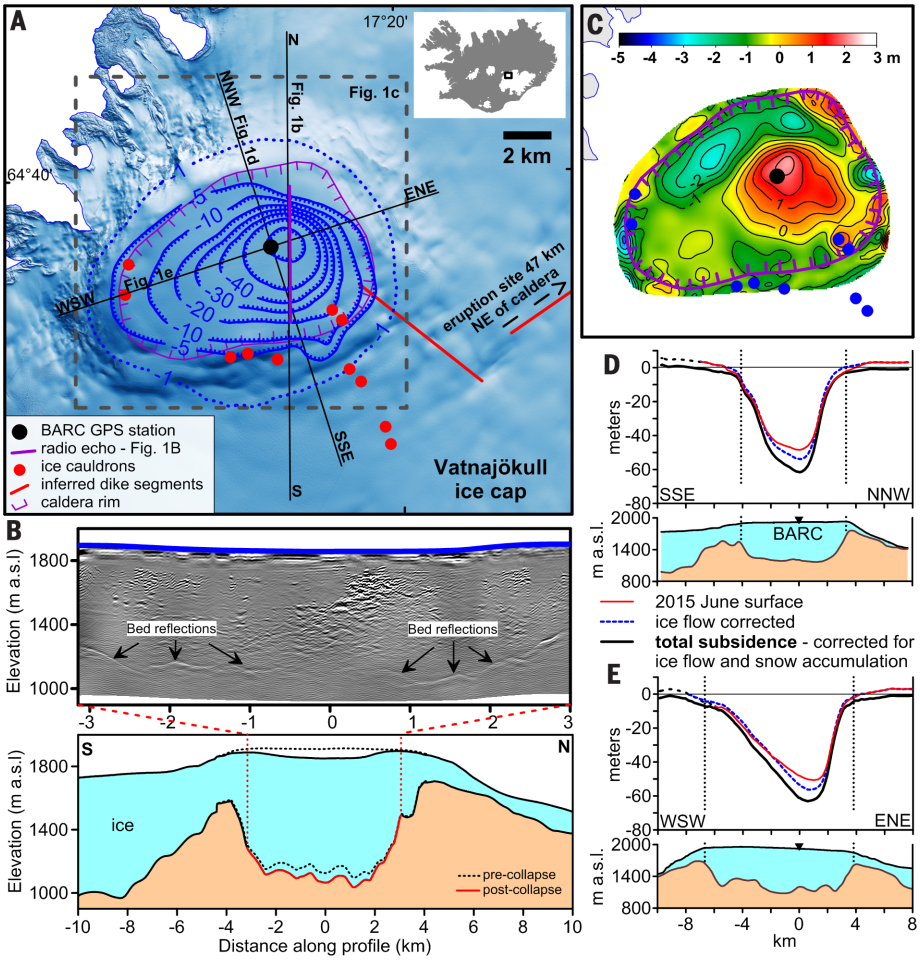



**Fig. 1. Bárdarbunga and geometry of collapse.** (**A**) Map showing the total caldera subsidence (in meters) at the end of collapse in February 2015. The blue dotted line is the 1-m subsidence contour. Minor sustained geothermal activity, monitored from aircraft, increased during the collapse, with pre-existing ice cauldrons deepening by up to 50 m and new ones forming at the southern margin and to the southeast of the caldera (*24*). (**B**) Radio-echo sounding profile from 3 February 2015 and a cross-section of the caldera showing the collapse. The precollapse topography is obtained by subtracting the subsidence observed at the surface. (**C**) Modeled changes in ice thickness at the end of February 2015 resulting from ice flow in response to caldera collapse (*24*). (**D**) North-northwest–south-southeast and (**E**) west-southwest–east-northeast cross-sections as measured in June 2015, corrected for winter snow accumulation in 2014–2015, measured in June 2015, and modeled vertical ice flow. Subsidence extends 2 to 3 km beyond the preexisting caldera rims vertical (dotted lines), at which it amounts to 3 to 11 m.

collapse resulted from a pressure drop in the reservoir as magma was laterally withdrawn into the propagating dike, with the latter possibly primarily driven by regional tectonic tensional stresses and topography (*20*).

The volume of the expanding dike on 20 August had reached ~0.25 km$^3$, increasing to 0.35 km$^3$ on 24 August (*20*), with the source of this magma being the reservoir beneath the caldera. The relatively minor caldera seismicity on 17 to 19 August indicates that the material overlying the magma reservoir deformed mostly elastically until it reached a critical failure point on 20 to 24 August. If we assume that the entire volume of eruptible magma within the reservoir was 1.9 ± 0.3 km$^3$, then the critical volume fraction required to reach the failure point and trigger the collapse (*23*) was 0.12 to 0.21.

## Ice flow, subsidence magnitude, and volume

As we recorded the caldera subsidence mainly on the ice (Fig. 1 and fig. S1), we made corrections and additional measurements to derive the underlying bedrock displacement. Our main data on ice surface changes and ice movements are repeated C-band radar altimeter surveys from aircraft, maps made from optical satellite images, and the continuously recording GPS station BARC (Bárdarbunga Caldera) that we set up in the center of the caldera on 13 September 2014. The observed velocities and displacements of the ice surface are displayed on Fig. 3, A and B. We use these observations to constrain 3D full-Stokes finite element modeling of ice flow in response to the collapse (*24*). The results show concentric flow, toward the point of maximum collapse within the caldera, with maximum ice thickening at the center of ~3 m by February 2015 (Fig. 1C and fig. S2). The maximum ice surface lowering of 62 ± 2 m, determined by aerial altimeter surveys, gives a maximum bedrock subsidence of 65 ± 3 m. Our data and models show that, apart from the concentric flow toward the deepest part of the subsidence (about 1 km east of BARC), horizontal flow was not much affected (Fig. 3A). We therefore conclude that suggestions of a large increase in ice flow out of the caldera during these events (*25*) cannot be fitted with our data.

Bedrock subsidence exceeding 1 m occurred within an area of 110 km$^2$ that extended beyond the preexisting caldera (Fig. 1 and fig. S1). After termination of collapse, the total subsidence at the preexisting caldera rims amounted to 3 to 11 m (Fig. 1, D and E). Using subglacial radio-echo soundings, we observed a down-sagged bedrock surface without any clear signs of fault offset (Fig. 1B) or indications of water bodies at the ice-bedrock interface. The limited resolution resulting from the 600- to 800-m ice thickness means that we cannot on the basis of the radio-echo results exclude the possibility of steep fault escarpments. However, substantial vertical fault movement at the base of the glacier would result in high strain rates within the basal ice, which would instantly fracture the ice fabric and propagate upward. During drainage of subglacial lakes in Iceland, large surface fractures induced by basal motion have been observed repeatedly (*26*) and can serve here as an analog for the possible surface manifestations of vertical basal motion. The absence of such surface ice fractures at Bárdarbunga indicates that no substantial fault escarpment formed at the bottom. The calculated collapse volume is 1.8 ± 0.2 km$^3$, not significantly different from the combined volume of erupted and intruded magma (Fig. 3B).

## Magma reservoir depth

Lava chemistry, surface gas composition, and geodetic modeling indicate drainage of a magma reservoir at a depth of ~12 km. The erupted lava is typical olivine tholeiite, with a relatively uniform chemical composition, consistent with efficient homogenization of melts before eruption. Several independent geobarometers (Fig. 4) yield an equilibrium pressure of 350 to 550 MPa, indicating that melt resided at depths of 11 to 16 km before the eruption. We obtained a similar result (14 ± 3 km) from analysis of subaerial gas measurements (Fig. 4). This depth concurs with our regional geodetic observations, which are dominated by a deflating source at 8- to 12-km depth beneath the Bárdarbunga caldera floor, after the reservoir cessation of dike-related deformation in mid-September (figs. S3 and S4).

## Seismicity and subsurface structure

We used seismic data and distinct element method (DEM) numerical modeling (*24*) to characterize the deeper collapse structure as the reactivation of a steeply inclined ring fault (Fig. 5). We mostly

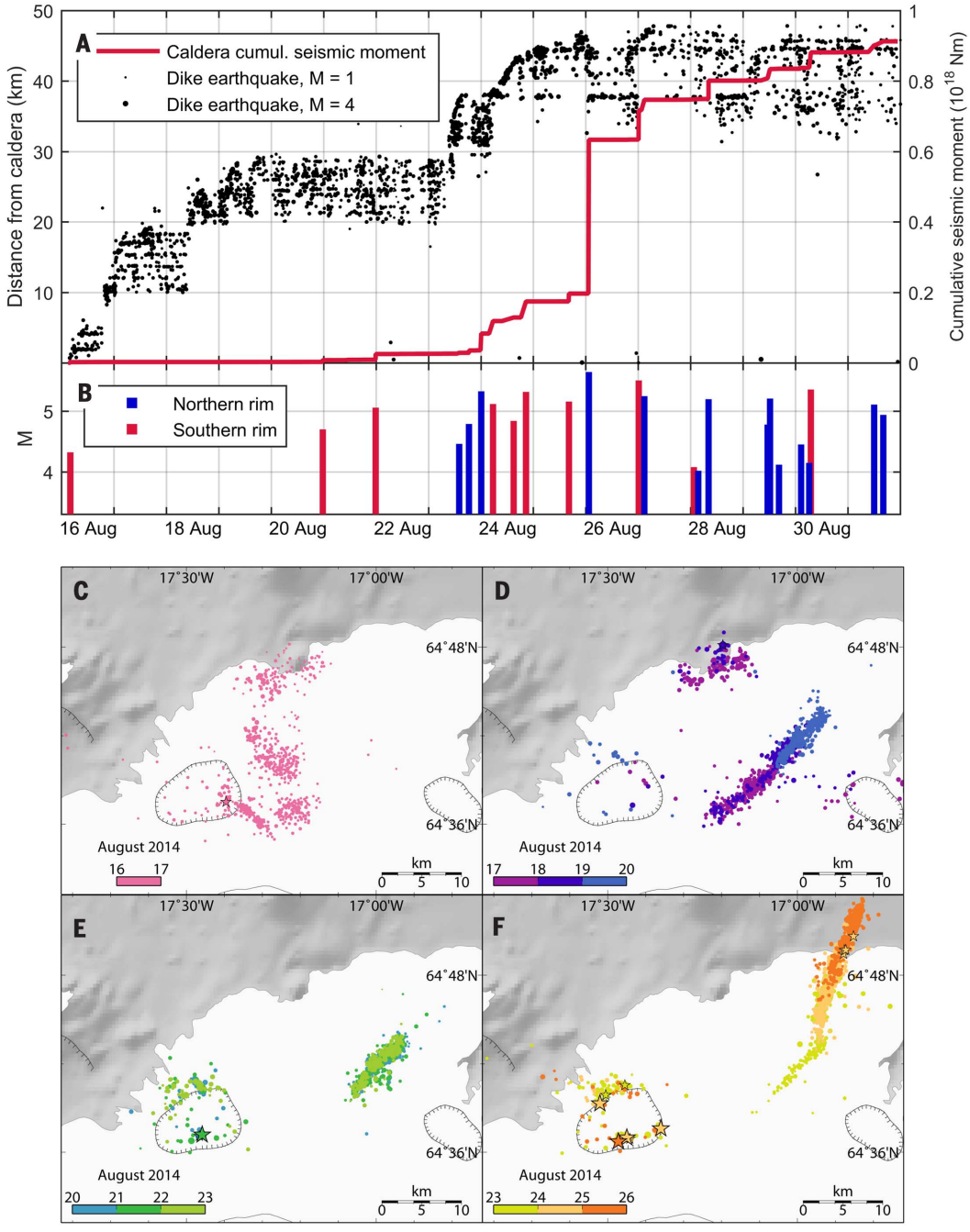



**Fig. 2. Onset of caldera collapse.** (**A**) Cumulative seismic moment release from caldera earthquakes plotted together with distribution of seismicity along the dike length, using high-quality relative locations of earthquakes (*20*), for the time period when the dike progressed away from the caldera. Black dots show individual earthquakes, with dot size scaling with magnitude. (**B**) Significant caldera earthquakes with magnitudes above *M*4 plotted as impulses, where the height represents magnitude and color represents location on the southern or northern rim. (**C** to **F**) Map of northwest Vatnajökull showing earthquake epicenters on (C) 16 August, (**D**) 17 to 19 August, (**E**) 20 to 22 August, and (F) 23 to 25 August 2014. The stars indicate the larger than magnitude M4 earthquakes in the caldera.

observed seismicity at depths of 0 to 9 km beneath the northern and southern caldera rims (Fig. 5B), with earthquakes being more numerous on the northern rim. This spatial pattern of seismicity is consistent with fracturing above a deflating magma reservoir that was elliptical in plan view and elongated east-northeast (*27*). In cross section, the hypocenters indicate a steeply (~80°) outward-dipping fault in the northern cluster, whereas in the southern cluster they indicate a vertical or near-vertical fault dip.

A series of DEM forward simulations of a magma reservoir and ring-fault system, as constrained by the hypocenter distribution and by the geobarometry data, tested the above structural interpretation against the observed north-northwest–south-southeast subsidence profile. The models indicate that a preexisting and relatively low friction (coefficient of 0.1 to 0.2) ring-fault system controlled the subsidence at depth (Fig. 5, C and D). Our best-fitting models had preexisting faults dipping at 80° to 85° away from the caldera center on the north side and at 85° to 90° toward the caldera center on the south side. The modeled preexisting faults lay at 1 to 2 km below the surface on the north side and 3 to 4 km on the south side. Modeling of a more complex fault geometry or the inclusion of greater material heterogeneity may further improve the data fit but presently lacks robust geophysical constraints. The arrangement of an outward-dipping fault on one side of a caldera and an inward-dipping fault on the other is typical of "asymmetric" or "trapdoor-like" collapses produced in past analog and numerical modeling studies (*8*, *28*, *29*). It also occurs at Glencoe (*29*) and Tendurek (*30*) volcanoes. Finally, our finding is consistent with past seismological results that defined a very similar ring-fault geometry during the last period of activity at Bárdarbunga in 1996 (*31*).

Through regional moment tensor (MT) inversion, we infer that the source mechanisms of 77

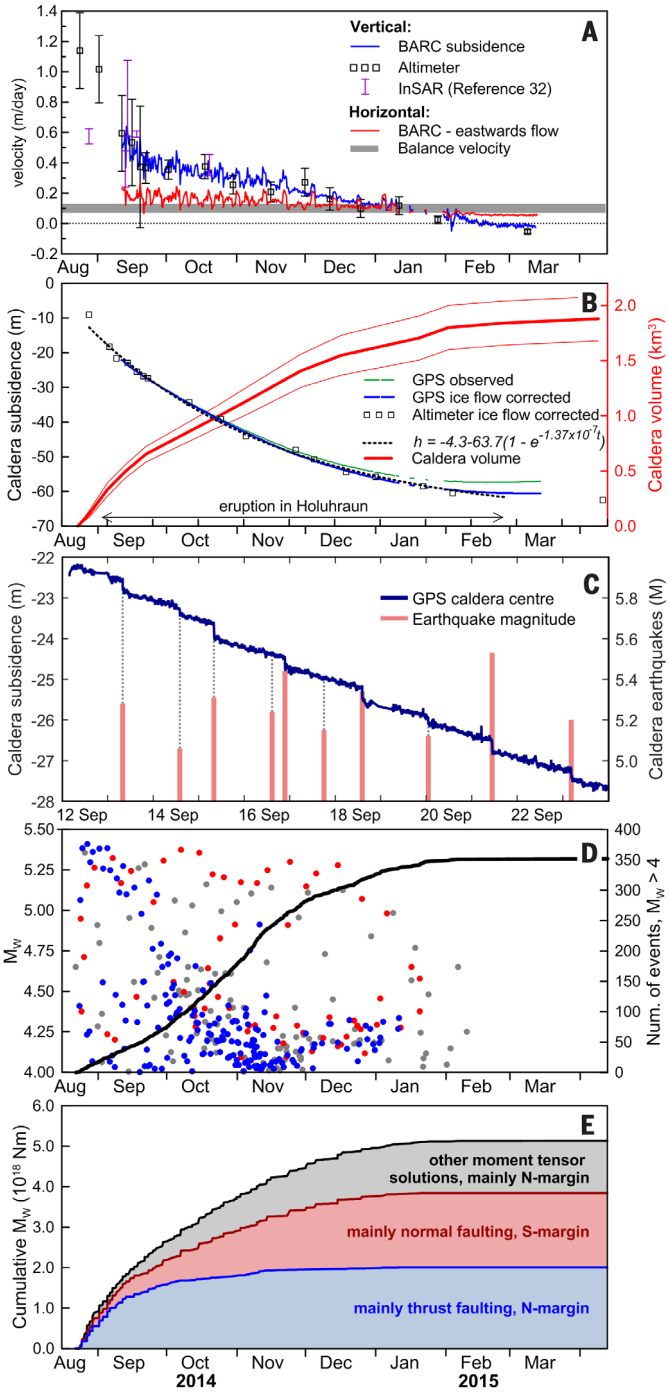



**Fig. 3. Time series of collapse.** (**A**) Vertical and horizontal velocities measured at the BARC GPS station in the center of the caldera (Fig. 1A), including the rate of vertical ice surface subsidence found from altimeter aircraft data and optical satellite photogrammetry. The horizontal balance velocity is obtained by estimating the rate of the eastward ice flow required to transport the net snow accumulation of the glacier to the west of BARC. InSAR-derived vertical velocities are based on (*32*). (**B**) Subsidence at the center of the caldera and subsidence volume evolution. The volume of the subsidence is obtained by subtracting the mapped surface from the precollapse surface. The caldera subsidence curve is fitted with an equation of the same form as Eq. 1. (**C**) High-resolution GPS for 12 to 23 September, showing $M > 5$ earthquakes coinciding with 20- to 40-cm rapid collapse, superimposed on gradual subsidence. The size of the steps is influenced by the location of BARC relative to the earthquake centroids and cannot be used directly to infer the proportion of ring-fault slip that ruptured seismically or aseismically. (**D**) Cumulative number of $M > 4$ caldera earthquakes, with magnitude evolution colored in red, blue, and gray representing clusters on the southern rim, the northern rim, and smaller clusters, respectively (see fig. S5). (**E**) Cumulative seismic moment for $M > 4$ caldera earthquakes.

$M > 5$ events (fig. S5) confined to two clusters beneath the northern and southern rim regions show contributions of both shear and nonshear components. The shear components indicate possible ruptures of segments on the ring fault. Shear failure on inward-dipping ring faults, or the sudden injection of magma in horizontal fissures forming sills, have been proposed (*32*) to explain the shear components of the observed earthquakes at Bárdarbunga. We, however, narrowed down on plausible solutions by using the micro-earthquakes (Fig. 5A). The moment tensor solutions are well constrained, but the inferred dip of the shear plane that we obtain is uncertain since the nonshear component, in this case a negative, subvertical compensated linear vector dipole (vCLVD), is dominant. As a result, the shear orientation obtained depends very much on the decomposition approach.

By using the constraint of the steeply outward-dipping ring fault on the northern cluster, we derive an MT solution that is a combination of a negative vCLVD and steep east-west–striking reverse faulting (Fig. 3, D and E, and fig. S5). In contrast, standard decomposition of the northern cluster MTs provides normal faulting along steep north-south–striking planes, a result that is inconsistent with the observed main fault orientation. The southern cluster MTs are consistent with being composed of families (*33*) of steep normal-faulting earthquakes.

The large, negative vCLVD indicates a combination of downward contraction and horizontal expansion, as has been observed in mines as well as in volcanic calderas during collapses [e.g., (*31, 34*)]. This could imply failure of support structures directly above or even within the magmatic reservoir or the sudden response of the reservoir fluid to vertical compression.

## Temporal development of subsidence and related seismicity

Subsidence occurred gradually during the eruption (Fig. 3B). From an initial rate in the caldera center of ~1 m/day during the first 20 days (Fig. 3B), subsidence declined in a near exponential manner with time (*24*). Subsidence terminated when the eruption ended in February 2015. We can associate some of the $M > 5$ caldera earthquakes, during the first couple of months of activity, with drops of 10 to 40 cm, but subsidence was otherwise continuous (Fig. 3C). The gradual decline in the rates of subsidence and caldera volume growth is mirrored by a decline in the cumulative seismic moment, the latter reflecting a decrease in the number of larger earthquakes with time (Fig. 3, D and E). Nonetheless, in terms of the cumulative seismic moment of $5.07 \times 10^{18}$ Nm for the $M > 4.0$ events, this collapse is the second-largest recorded, after that of Katmai (1912) (*35*). The geodetic moment depends on the shear modulus, the fault area, and the amount of slip assumed. The shear modulus could be very low in regions of intense faulting, such as on a caldera ring fault. The possible range of the geodetic moment is found by considering a ring fault reaching from the surface to 12-km depth, 60 m of slip, and a shear modulus over a wide range, 2 to 20 GPa. This results in a moment of $4 \times 10^{19}$ to $4 \times 10^{20}$ Nm, or 10 to 100 times the cumulative seismic moment of the earthquakes. This difference is consistent with the modeling of surface deformation observed during one of the events (fig. S7).

## Caldera-flow path interaction and piston collapse modeling

We see a short-term (multihourly) mechanical coupling of the collapsing caldera and the distal dike (south of eruption site) in the timing of earthquakes in the dike and at the caldera (Fig. 6A). Within a 6-hour window before and after large caldera earthquakes, the frequency of dike earthquakes was increased relative to background rate (*24*). We observed this pattern in the data after the beginning of October 2014, when the dike had stopped propagating and a quasisteady magma flow path had developed, until February

2015 when seismic activity stopped. For the 3 hours after caldera earthquakes with magnitude $M > 4.6$, as well as for the 3 hours before caldera earthquakes with $M > 4.0$, the increase in seismicity was significant (*24*) ($P = 0.05$) (Fig. 6 and fig. S8).

At Bárdarbunga, communication therefore existed between caldera subsidence events and pressure changes in a conduit up to 48 km away. Spatiotemporal patterns of tilt at Kilauea volcano, Hawaii, show a similar phenomenon that can be explained by the propagation of pressure transients within an elastically deformable dike (*36*). By analogy, we can make the interpretation that caldera earthquakes may generate a pressure pulse that leads to increased seismicity at the end of the dike. The communication could be two-way, although it is difficult to explain a pressure pulse from the dike toward the caldera. One possibility is that readjustment of the dike (e.g., sudden unblocking) can increase the dike volume slightly and subsequently lower the magma pressure, which then translates back to the caldera. The communication may also be entirely one-way, from the caldera to the dike: Smaller caldera earthquakes, and/or aseismic deformation at depth just above the magma reservoir, may precede a large caldera earthquake, increasing dike pressure and dike seismicity.

We explain the longer-term (weeks to months scale) coupling in the form of the gradually declining rates of caldera subsidence, caldera volume change, and lava eruption (Figs. 3B and 6B) with a model of a collapsing piston overlying a pressurized magma reservoir. We assume that the reservoir pressure and fault friction each partially support the piston weight (*24*). Drainage of magma reduces the reservoir pressure and causes piston subsidence (fig. S6). This in turn raises the reservoir pressure, leading to a feedback loop that maintains quasiconstant pressure at the magma reservoir top and drives further magma drainage. The pressure feeding the eruption drops, however, due to the reduction in the hydraulic head of magma over time. Kumagai *et al.* (*37*) also used a piston model to explain caldera collapse at Miyakejima in 2000, but in their model no change in hydraulic head was assumed and outflow rate was held constant.

Assuming that the time-averaged resistive force due to friction on the ring faults remains constant and that magma flow is laminar through a cylindrical pipe with radius $r$ and conduit length $L$ with $L >> r$, then

$$\Delta P \approx \Delta P_0 \exp\left[-\frac{\pi \rho g r^4}{8 A \eta L} t\right] \quad (1)$$

where $\Delta P$ is the driving overpressure, $\Delta P_0$ is the initial driving overpressure, $\rho$ is the density of the magma, $g$ is gravitational acceleration, $A$ is the cross-sectional area of the magma reservoir, $\eta$ is the dynamic viscosity of the magma, and $t$ is time (*24*). We estimated $\Delta P_0$ and the constant in the exponent, assuming that the measured subsidence within the caldera represents the decrease in magma reservoir height with time (Fig. 3B). This represents a minimum estimate for $\Delta P_0$, because there may also have been dilation at depth. The model also fits the measured caldera volume change (Fig. 3B) and eruption rate (Fig. 6B). This model predicts the same form of decay in flow

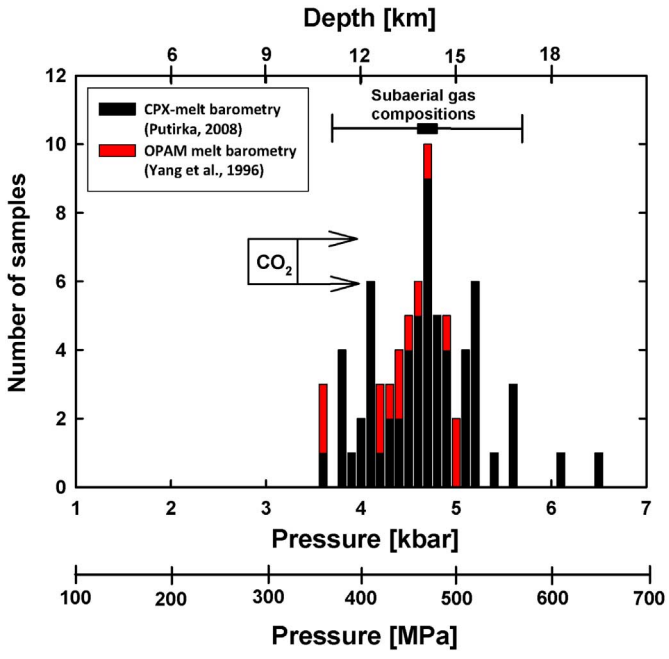



**Fig. 4. Magma reservoir depth from geobarometric and subaerial gas analysis.** The vertical shaded-bar crossing the graph, indicates the minimum pressure obtained (~300 MPa) from density barometry of plagioclase hosted $CO_2$-bearing fluid inclusions. Due to post-entrapment modifications of fluid inclusions during magma ascent, this value is likely an under-estimate, as indicated by the dashed arrows. The results from the analysis of subaerial gas compositions are based on FTIR and MultiGAS measurements (*24*).

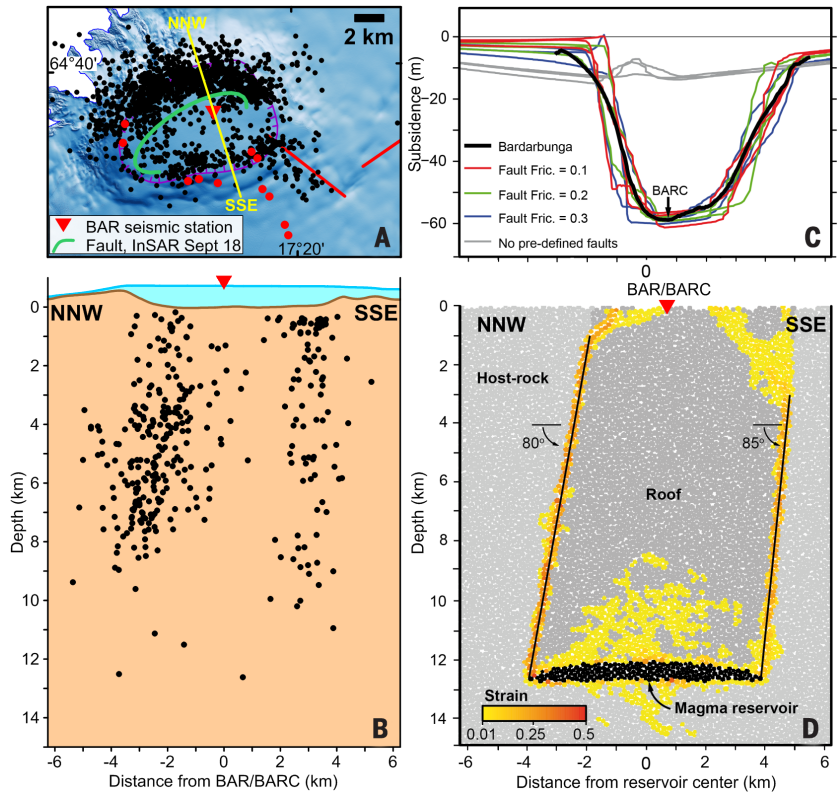



**Fig. 5. Fault geometry and collapse modeling.** (**A**) Earthquakes 1 August to 17 October 2014. (**B**) Seismicity along a 2- to 4-km-wide strip on the north-northwest–south-southeast cross section; depth relative to bedrock caldera floor. (**C** and **D**) Two-dimensional DEM modeling of the collapse, constrained by subsidence geometry, earthquake locations in (B), and the geobarometry (Fig. 4). The geometry illustrated in (D) obtained the best agreement with the observations. The color scale shows the maximum finite shear strain. Surface displacement profiles for different preexisting fault frictions are provided in (C). Three model realizations are shown for each friction value.

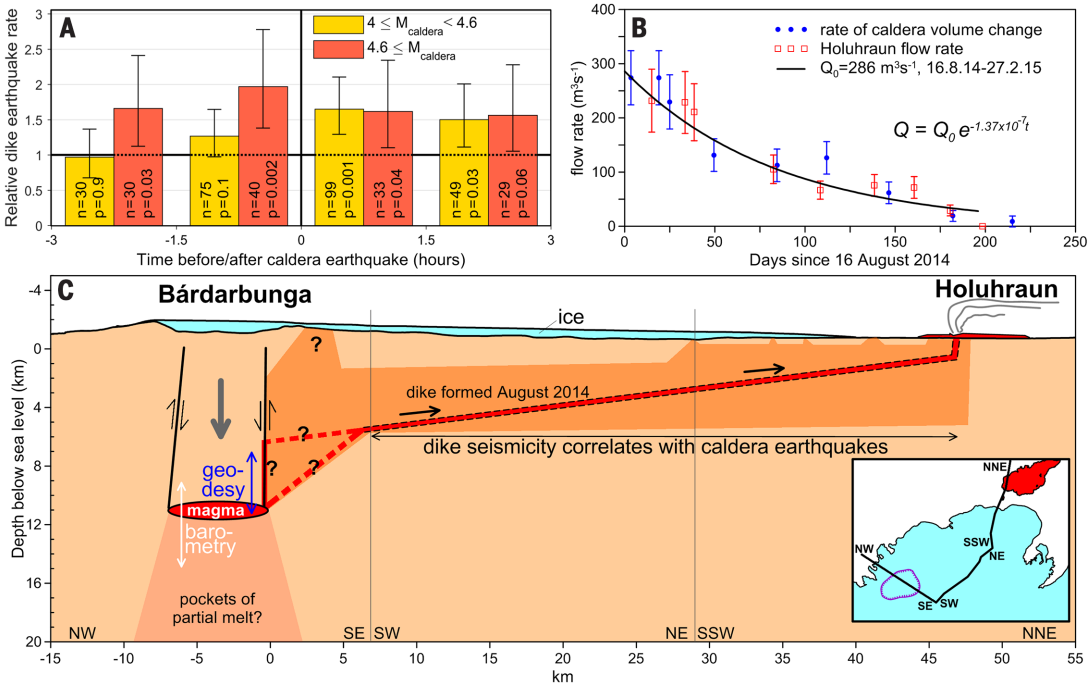



**Fig. 6. Caldera-magma flow path interaction.** (**A**) Rate of dike earthquakes relative to background levels before and after caldera earthquakes of magnitude > *M*4. The *P* values indicate the two-sided significance, and *n* is the number of earthquakes used. Error bars indicate 90% confidence intervals (*24*). (**B**) Exponential model of magma flow rate constrained by caldera GPS subsidence (*24*) compared with rate of volume change in caldera and eruption rate in Holuhraun. The eruption stopped on day 194 (27 February 2015) before the driving pressure reached zero, as expected if the conduit becomes clogged by solidifying magma as the flow rate drops. Caldera subsidence also terminated around the end of February (see GPS ice-flow corrected subsidence in Fig. 3B). (**C**) Schematic cross section of caldera–magma reservoir–pipe-like magma flow path and eruption site after dike formation (*20*, *21*). The inferred magma reservoir is set at 12 km below bedrock caldera floor. It is possible that magma ascended first along the ring fault before forming the dike above 6- to 10-km depth. We indicate the constraints on depth to magma reservoir from geobarometry with a white arrow and from geodesy with a blue arrow.

rate (exponential) as the standard Wadge model of depressurization of an overpressured magma body (*38*) but by a different mechanism. The feedback mechanism of repressurization from the ongoing piston collapse enhanced the length and speed of dike propagation and the duration of the eruption. In this model, therefore, both the eruption drives the collapse and collapse drives the eruption.

## Overview and implications

Table 1 contextualizes the key features of the 2014–2015 Bárdarbunga collapse with respect to those of the six other collapses instrumentally monitored to date. The areal extent of the Bárdarbunga collapse (110 $km^2$) is the largest yet observed historically and is comparable to that associated with major silicic eruptions in the geological record (*6*). The total subsidence (65 m) is one to two orders of magnitude smaller than all past collapses listed here, but the large area means that it has the fourth-largest collapse volume (1.8 $km^3$) overall. The erupted volume (1.4 $km^3$) is the largest of the observed mafic collapses so far, although considerable uncertainty surrounds the volumes associated with the collapse of Fernandina. In volume terms, both the silicic eruptions and collapses of Katmai and Pinatubo were two to six times as large. The cumulative seismic energy release at Bárdarbunga ($25 \times 10^{13}$ J) (see Table 1) is dwarfed by that of Katmai ($1600 \times 10^{13}$ J) but is similar to Miyakejima ($22 \times 10^{13}$ J), despite the much smaller area of the latter (1.9 $km^2$). This is explained by the much greater subsidence at Katmai (>1200 m) and Miyakejima (>1600 m). The gradual collapse of Bárdarbunga had the second-longest duration (190 days) yet recorded. Only the duration of collapse at Tolbachik (515 days) exceeds it. Finally, Bárdarbunga has the longest confirmed length of an associated lateral intrusion (48 km) and the longest distance to the main vent (40 km).

Our data and modeling show that withdrawal and eruption of magma triggered the collapse at Bárdarbunga. For the likely depth-to-diameter ratio of the magma reservoir, the critical volume fraction required to trigger the onset of collapse (0.12 to 0.21) was much lower than that predicted by past analytical and analog modeling (*23*, *39*). A similar inference of low critical volume fractions at La Reunion and Miyakejima (*16*) was explained as a consequence of the reactivation of preexisting ring faults, a proposition in line with our observations and analysis of the Bárdarbunga collapse.

Nonetheless, we also show that there is a tight mechanical interplay between collapse and eruption throughout the process once collapse has started, with eruption driving collapse, and vice versa, on both hourly and eruption-long time scales. For the longer time-scale coupling, the results also show that the physical properties of both the magma reservoir roof and the magma pathway regulate caldera collapse and magma outflow rate. Consequently, collapse at Bárdarbunga occurred gradually and at a steadily (exponentially) declining rate. This is a very similar pattern to that inferred for the 1968 Fernandina collapse (*2*, *16*). In contrast to some model predictions (*40*) and to the 2007 collapse of Piton de la Fournaise (*16*), we found no evidence for rapid and sustained pressure increase in the magma reservoir as a result of collapse, possibly due to substantial ductile behavior of the roof of the larger and deeper Bárdarbunga magma reservoir (*13*, *16*).

The question of whether or to what extent our understanding of caldera collapse at mafic volcanoes such as Bárdarbunga is transferrable to large silicic systems remains an open one. On the one hand, the gradual nature of collapse at Bárdarbunga and Fernandina contrasts with the highly punctuated collapse style inferred during explosive silicic eruptions like Katmai and Pinatubo (*2*, *41*). In addition, collapse at silicic volcanoes is generally considered to be triggered by eruption through a central vent rather than through the lateral withdrawal mechanism seen at Bárdarbunga. On the other hand, the most silicic of all instrumentally monitored collapses, Katmai, was also clearly associated with a lateral withdrawal. This mechanism could therefore be more widespread at silicic calderas than commonly considered. Furthermore, the punctuated collapse style inferred for the Katmai and Pinatubo cases rests on large (M5-7) earthquakes that yield dramatic steps in a curve of cumulative seismic release against time (2). The locations and source mechanisms of these apparently collapse-related earthquakes are poorly constrained, however. As best highlighted in the case of Pinatubo, a regional tectonic origin for them cannot be precluded (*15*, *35*, *42*, *43*). Consequently, Bárdarbunga

**Table 1. Instrumentally monitored caldera collapses since 1900 CE.** References for data: Bárdarbunga (*20*) and this study; La Reunion (*12*, *16*, *44*–*46*); Miyakejima (*2*, *11*, *34*, *47*–*49*); Tolbachik (*50*–*53*); Fernandina (*2*, *9*, *54*); Katmai (*35*, *41*, *55*); and Pinatubo (*2*, *15*, *42*, *43*, *56*).

| Volcano | Year | Magma | Maximum subsidence (m) | Collapse duration (days) | Collapse area ($km^2$) | Collapse volume ($km^3$) | Reservoir depth (km) | Intrusion volume ($km^3$)† | Erupted volume ($km^3$)‡ | Total magma volume ($km^3$) | Distance to vent (km)§ | Intrusion length (km)‖ | Seismic energy ($\times 10^{13}$ J)¶ | Max. EQ# |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bárdarbunga | 2015 | Basalt | ~65 | 190 | 110 | 1.8 | 11–16 | 0.5 | 1.4 | 1.9 | 40 | 48 | 25 | 5.8 |
| La Reunion | 2007 | Basalt | ~450 | 2 | 0.82 | 0.1 | 2–3 | 0.02 | 0.14 | 0.16 | 7 | 7 | ? | 3.2 |
| Miyakejima | 2000 | Basalt | ~1600 | 40 | 1.9 | 0.6 | 4–7 | 1.2 | 0.01 | 1.21 | 5.6 | 35 | 22 | 5.6 |
| Tolbachik | 1976 | Basalt | >500 | 515 | 2.5 | 0.35 | 4–6? | ? | 1.2 | >1.2 | 28 | ~45 | ? | 2.9 |
| Fernandina | 1969 | Basalt | ~350 | 12 | 7 | 2 | ? | ? | 0.2 | >0.2 | 10.5? | 10.5? | 2 | 5.2 |
| Katmai | 1912 | Rhyolite | >1300 | 3 | 8.8 | 5.5 | 2–5 | ? | *13.5* | *>13.5* | 10 | 10 | *1600* | *7* |
| Pinatubo* | 1991 | Dacite | *~900* | *2* | *4* | *2* | *7–11* | ? | *4.5* | *4.5* | *1* | *4* | *2* | *5.7* |

*All caldera collapses except Pinatubo formed in association with lateral withdrawal and intrusion of magma. †Intrusion volume values are typically constrained by inversions of data from geodetic networks and so are available only for the most recent events. ‡Erupted volumes are given as dense-rock equivalent—i.e., with porosity removed. §Distance measured from center of caldera to the most distant known vent active during collapse. ‖Estimated horizontal length of the intrusion, from locations of seismicity and/or inversions of geodetic data in all cases except Katmai. For Katmai and Fernandina, intrusion length is estimated as the distance from caldera to vent and is hence a minimum value. ¶Cumulative seismic energy release calculated by converting the cumulative scalar moments ($M_0$) by using a factor of $5 \times 10^{-5}$ [from an energy-moment relationship determined by Kanamori *et al*. (*57*)]. #Maximum earthquake magnitude (EQ) associated with caldera formation. Magnitude determined from surface waves ($M_s$) is given for Tolbachik (*53*), Katmai (*35*), and Fernandina (*54*). For La Reunion, Earthquake duration magnitude ($M_d$) is used (*12*, *44*). For Miyakejima and Bárdarbunga, the maximum moment magnitude ($M_w$) for collapse-related very-long-period events is given (*34*, *58*, this study).

2014–2015 provides our clearest picture yet of how caldera collapse can be triggered during large eruptions and how the dynamics of the subterranean magma flow path and the interaction with magma reservoir pressure regulates eruption rates and the rate of collapse.

## Materials and methods

We mapped changes in ice surface topography with time by using aircraft-based radar altimetry, satellite optical photogrammetry, and a continuously recording GPS station installed on the ice surface. We modeled the ice flow within the caldera by using a full-Stokes finite element model constrained by the horizontal GPS velocities. We measured the caldera bedrock topography through radio-echo sounding. We mapped changes in lava volume with a ground-based theodolite and aircraft altimetry. Bulk density of the lava was determined from gravity profiles. We used interferometric synthetic aperture radar (InSAR) to determine coseismic ground deformation at the caldera for several 24-hour periods spanning large caldera earthquakes. We ran elastic dislocation models of this deformation to infer the location of faults that slipped during the earthquakes. Microearthquakes ($M \leq 2$) at the caldera were recorded by the Icelandic national seismic network and relatively relocated by using a standard 1D velocity model. The microearthquakes were shifted southward by using the faults inferred from the InSAR modeling. We performed a regional moment tensor inversion for all events with $M > 5$, adopting a full (MT point source approximation and elucidating double-couple, compensated linear vector dipole and isotropic components. We extended the interpretation of source processes to smaller events by applying a waveform similarity analysis, and we more precisely determined locations by using relative location techniques. We evaluated the role of preexisting ring-fault structures on collapse by using 2D DEM modeling. We determined the approximate depth of the magma reservoir from a point-pressure source model in an elastic half space, constrained by post-rifting InSAR and GPS data. We computed caldera-dike seismicity correlations by calculating the number of dike earthquakes in 1.5-hour bins before and after large caldera earthquakes; $P$ values were computed with a likelihood ratio test. We analyzed major element compositions of minerals and glasses by using an electron microprobe. Fluid inclusions within phenocrysts were analyzed by optical microscopy and confocal Raman spectroscopy. Three independent thermobarometers constrained the depth of magma accumulation before the onset of the eruption: glass thermobarometry, clinopyroxene-liquid thermobarometry, and $CO_2$ density barometry. We also measured the composition of subaerial eruptive gases on multiple occasions with an open-path Fourier transform infrared spectrometer (FTIR) and a multicomponent gas analyzer system (MultiGAS). Finally, we modeled caldera subsidence and eruption rate by considering analytically a collapsing piston overlying a pressurized magma reservoir and driving magma flow through a cylindrical pipe.

### REFERENCES AND NOTES


1. M. Branney, V. Acocella, in *The Encyclopedia of Volcanoes*, H. Sigurdsson, B. Houghton, H. Rymer, Eds. (Academic Press, Academic Press, Heidelberg, Amsterdam, Boston, ed. 2, 2015), pp. 229–315.
2. J. Stix, T. Kobayashi, Magma dynamics and collapse mechanisms during four historic caldera-forming events. *J. Geophys. Res.* **113**, B09205 (2008). doi: 10.1029/2007JB005073
3. A. R. McBirney, A historical note on the origin of calderas. *J. Volcanol. Geotherm. Res.* **42**, 303–306 (1990). doi: 10.1016/0377-0273(90)90006-2
4. R. S. J. Sparks *et al*., Ignimbrites of the Cerro Galán caldera, NW Argentina. *J. Volcanol. Geotherm. Res.* **24**, 205–248 (1985). doi: 10.1016/0377-0273(85)90071-X
5. T. H. Druitt, R. S. J. Sparks, On the formation of calderas during ignimbrite eruptions. *Nature* **310**, 679–681 (1984). doi: 10.1038/310679a0
6. P. W. Lipman, Subsidence of ash-flow calderas: Relation to caldera size and magma-chamber geometry. *Bull. Volc* **59**, 198–218 (1997). doi: 10.1007/s004450050186
7. V. Acocella, Understanding caldera structure and development: An overview of analogue models compared to natural calderas. *Earth Sci. Rev.* **85**, 125–160 (2007). doi: 10.1016/j.earscirev.2007.08.004
8. O. Roche, T. H. Druitt, O. Merle, Experimental study of caldera formation. *J. Geophys. Res.* **105**, 395–416 (2000). doi: 10.1029/1999JB900298
9. T. Simkin, K. A. Howard, Caldera collapse in the Galapagos Islands, 1968: The largest known collapse since 1912 followed a flank eruption and explosive volcanism within the caldera. *Science* **169**, 429–437 (1970). doi: 10.1126/science.169.3944.429; pmid: 17739001
10. S. A. Fedotov, Y. K. Markhinin, *The Great Tolbachik Fissure Eruption. Geological and Geophysical Data 1975-1976* (Cambridge Univ. Press, Cambridge, 2011).
11. N. Geshi, T. Shimano, S. Chiba, S. Nakada, Caldera collapse during the 2000 eruption of Miyakejima Volcano, Japan. *Bull. Volcanol.* **64**, 55–68 (2002). doi: 10.1007/s00445-001-0184-z
12. T. Staudacher *et al*., The April 2007 eruption and the Dolomieu crater collapse, two major events at Piton de la Fournaise (La Réunion Island, Indian Ocean). *J. Volcanol. Geotherm. Res.* **184**, 126–137 (2009). doi: 10.1016/j.jvolgeores.2008.11.005
13. E. P. Holohan, M. P. J. Schöpfer, J. J. Walsh, Stress evolution during caldera collapse. *Earth Planet. Sci. Lett.* **421**, 139–151 (2015). doi: 10.1016/j.epsl.2015.03.003
14. K. V. Cashman, G. Giordano, Calderas and magma reservoirs. *J. Volcanol. Geotherm. Res.* **288**, 28–45 (2014). doi: 10.1016/j.jvolgeores.2014.09.007
15. J. Battaglia, C. H. Thurber, J.-L. Got, C. A. Rowe, R. A. White, Precise relocation of earthquakes following the 15 June 1991 eruption of Mount Pinatubo (Philippines). *J. Geophys. Res.* **109**, B07302 (2004). doi: 10.1029/2003JB002959
16. L. Michon, F. Massin, V. Famin, V. Ferrazzini, G. Roult, Basaltic calderas: Collapse dynamics, edifice deformation, and variations of magma withdrawal. *J. Geophys. Res.* **116**, B03209 (2011). doi: 10.1029/2010JB007636
17. H. Björnsson, P. Einarsson, Volcanoes beneath Vatnajökull, Iceland: Evidence from radio-echo sounding, earthquakes and jökulhlaups. *Jokull* **40**, 147–148 (1990).
18. M. T. Gudmundsson, T. Högnadóttir, Volcanic systems and calderas in the Vatnajökull region, central Iceland, constraints on crustal structure from gravity data. *J. Geodyn.* **43**, 153–169 (2007). doi: 10.1016/j.jog.2006.09.015

19. T. Thordarson, G. Larsen, Volcanism in Iceland in historical time: Volcano types, eruption styles and eruption history. *J. Geodyn.* **43**, 118–152 (2007). doi: 10.1016/j.jog.2006.09.005
20. F. Sigmundsson *et al*., Segmented lateral dyke growth in a rifting event at Bárðarbunga volcanic system, Iceland. *Nature* **517**, 191–195 (2015). doi: 10.1038/nature14111; pmid: 25517098
21. T. Ágústsdóttir *et al*., Strike-slip faulting during the 2014 Bárðarbunga-Holuhraun dike intrusion, central Iceland. *Geophys. Res. Lett.* **43**, 1495–1503 (2016). 10.1002/2015GL067423
22. S. R. Gíslason *et al*., Environmental pressure from the 2014–15 eruption of Bárðarbunga volcano, Iceland. *Geochem. Perspect. Lett* **1**, 84–93 (2015). doi: 10.7185/geochemlet.1509
23. O. Roche, T. H. Druitt, Onset of caldera collapse during ignimbrite eruptions. *Earth Planet. Sci. Lett.* **191**, 191–202 (2001). doi: 10.1016/S0012-821X(01)00428-9
24. Materials and methods are available as supplementary materials on *Science* Online.
25. J. Browning, A. Gudmundsson, Surface displacements resulting from magma-chamber roof subsidence, with application to the 2014-2015 Bardarbunga-Holuhraun volcanotectonic episode in Iceland. *J. Volcanol. Geotherm. Res.* **308**, 82–98 (2015). doi: 10.1016/j.jvolgeores.2015.10.015
26. H. Björnsson, *Hydrology of Ice Caps in Volcanic Regions, Rit. 45* (Soc. Sci. Isl., Reykjavík, 1988).
27. E. P. Holohan, V. R. Troll, B. Van Wyk de Vries, J. J. Walsh, T. R. Walter, Unzipping Long Valley: An explanation for vent migration patterns during an elliptical ring fracture eruption. *Geology* **36**, 323–326 (2008). doi: 10.1130/G24329A.1
28. E. P. Holohan *et al*., Origins of oblique-slip faulting during caldera subsidence. *J. Geophys. Res.* **118**, 1778–1794 (2013). doi: 10.1002/jgrb.50057
29. C. T. Clough, H. B. Maufe, E. B. Baley, The cauldron subsidence of Glen-Coe, and the associated igneous phenomena. *Q. J. Soc. Lond.* **65**, 611–678 (1909). doi: 10.1144/GSL.JGS.1909.065.01-04.35
30. H. Bathke, M. Nikkhoo, E. P. Holohan, T. R. Walter, Insights into the 3D architecture of an active caldera ring-fault at Tendürek volcano through modeling of geodetic data. *Earth Planet. Sci. Lett.* **422**, 157–168 (2015). doi: 10.1016/j.epsl.2015.03.041
31. A. Fichtner, H. Tkalcic, Insights into the kinematics of a volcanic caldera drop: Probabilistic finite-source inversion of the 1996 Bardarbunga, Iceland, earthquake. *Earth Planet. Sci. Lett.* **297**, 607–615 (2010). doi: 10.1016/j.epsl.2010.07.013
32. B. Riel *et al*., The collapse of Bárðarbunga Caldera, Iceland. *Geophys. J. Int.* **202**, 446–453 (2015). doi: 10.1093/gji/ggv157
33. S. Cesca, A. T. Sen, T. Dahm, Seismicity monitoring by cluster analysis of moment tensors. *Geophys. J. Int.* **196**, 1813–1826 (2014). doi: 10.1093/gji/ggt492
34. A. Shuler, G. Ekström, M. Nettles, Physical mechanisms for vertical-CLVD earthquakes at active volcanoes. *J. Geophys. Res.* **118**, 1569–1586 (2013). doi: 10.1002/jgrb.50131
35. K. Abe, Seismicity of the caldera-making eruption of Mount Katmai, Alaska in 1912. *Bull. Seismol. Soc. Am.* **82**, 175–191 (1992).
36. C. P. Montagna, H. M. Gonnermann, Magma flow between summit and Pu'u 'O'o at Kilauea Volcano, Hawaii. *Geochem. Geophys. Geosyst.* **14**, 2232–2246 (2013). doi: 10.1002/ggge.20145
37. H. Kumagai *et al*., Very-long-period seismic signals and caldera formation at Miyake Island, Japan. *Sci.* **293**, 687–690 (2001). doi: 10.1126/science.1062136; pmid: 11474109
38. G. Wadge, Steady state volcanism: Evidence from eruption histories of polygenetic volcanoes. *J. Geophys. Res.* **87**, 4035–4049 (1982). doi: 10.1029/JB087iB05p04035
39. A. Geyer, A. Folch, J. Martí, Relationship between caldera collapse and magma chamber withdrawal: An experimental approach. *J. Volcanol. Geotherm. Res.* **157**, 375–386 (2006). doi: 10.1016/j.jvolgeores.2006.05.001
40. A. Folch, J. Marti, Time-dependent chamber and vent conditions during explosive caldera-forming eruptions. *Earth Planet. Sci. Lett.* **280**, 246–253 (2009). doi: 10.1016/j.epsl.2009.01.035
41. W. Hildreth, J. Fierstein, Katmai volcanic cluster and the great eruption of 1912. *Geophys. J. Int.* **112**, 446–453 (2000).
42. B. C. Bautista *et al*., in *Fire and Mud: Eruptions and Lahars of Mount Pinatubo, Philippines*, C. G. Newhall, R. S. Punongbayan, Eds. (Philippine Institute of Volcanology and Seismology & University of Washington Press, Quezon City, Seattle & London, 1996), pp. 351–370.
43. J. Mori *et al*., in *Fire and Mud: Eruptions and Lahars of Mount Pinatubo, Philippines*, C. G. Newhall, R. S. Punongbayan, Eds. (Philippine Institute of Volcanology and Seismology & University of Washington Press, Quezon City, Seattle & London, 1996), pp. 339–350.
44. L. Michon, T. Staudacher, V. Ferrazzini, P. Bachèlery, J. Marti, April 2007 collapse of Piton de la Fournaise: A new example of caldera formation. *Geophys. Res. Lett.* **34**, L21301 (2007). doi: 10.1029/2007GL031248
45. F. R. Fontaine, G. Roult, L. Michon, G. Barruol, A. Di Muro, The 2007 eruptions and caldera collapse of the Piton de la Fournaise volcano (La Reunion Island) from tilt analysis at a single very broadband seismic station. *Geophys. Res. Lett.* **41**, 2803–2811 (2014). doi: 10.1002/2014GL059691
46. A. Di Muro *et al*., The shallow plumbing system of Piton de la Fournaise Volcano (La Reunion Island, Indian Ocean) revealed by the major 2007 caldera-forming eruption. *J. Petrol.* **55**, 1287–1315 (2014). doi: 10.1093/petrology/egu025
47. T. Kaneko *et al*., Submarine flank eruption preceding caldera subsidence during the 2000 eruption of Miyakejima Volcano, Japan. *Bull. Volcanol.* **67**, 243–253 (2005). doi: 10.1007/s00445-004-0407-1
48. M. Amma-Miyasaka, M. Nakagawa, S. Nakada, Magma plumbing system of the 2000 eruption of Miyakejima Volcano, Japan. *Bull. Volcanol.* **67**, 254–267 (2005). doi: 10.1007/s00445-004-0408-0
49. T. Nishimura *et al*., Crustal deformation caused by magma migration in the northern Izu Islands, Japan. *Geophys. Res. Lett.* **28**, 3745–3748 (2001). doi: 10.1029/2001GL013051
50. S. A. Fedotov, L. B. Slavina, S. L. Senyukov, M. S. Kuchay, Seismic processes and migration of magma during the Great Tolbachik Fissure Eruption of 1975-1976 and Tolbachik Fissure Eruption of 2012-2013, Kamchatka Peninsula. *Izv. Atmos. Ocean. Phys.* **51**, 667–687 (2015). doi: 10.1134/S000143381507004X
51. S. A. Fedotov, I. S. Utkin, L. I. Utkina, The peripheral magma chamber of Ploskii Tolbachik, a Kamchatka basaltic volcano: Activity, location and depth, dimensions, and their changes based on magma discharge observations. *J. Volcanol. Seismol.* **5**, 369–385 (2011). doi: 10.1134/S0742046311060042
52. P. Doubik, B. E. Hill, Magmatic and hydromagmatic conduit development during the 1975 Tolbachik Eruption, Kamchatka, with implications for hazards assessment at Yucca Mountain, NV. *J. Volcanol. Geotherm. Res.* **91**, 43–64 (1999). doi: 10.1016/S0377-0273(99)00052-9
53. S. A. Fedotov, A. M. Chirkov, N. A. Gusev, G. N. Kovalev, Y. B. Slezin, The large fissure eruption in the region of Plosky Tolbachik volcano in Kamchatka, 1975–1976. *Bull. Volcanol.* **43**, 47–60 (1980). doi: 10.1007/BF02597610
54. J. Filson, T. Simkin, L. K. Leu, Seismicity of a caldera collapse: Galapagos Islands 1968. *J. Geophys. Res.* **78**, 8591–8622 (1973). doi: 10.1029/JB078i035p08591
55. J. E. Hammer, M. J. Rutherford, W. Hildreth, Magma storage prior to the 1912 eruption at Novarupta, Alaska. *Contrib. Mineral. Petrol.* **144**, 144–162 (2002). doi: 10.1007/s00410-002-0393-2
56. M. J. Rutherford, J. D. Devine, in *Fire and Mud: Eruptions and Lahars of Mount Pinatubo, Philippines*, C. G. Newhall, R. S. Punongbayan, Eds. (Philippine Institute of Volcanology and Seismology & University of Washington Press, Quezon City, Seattle & London, 1996), pp. 751–766.
57. H. Kanamori *et al*., Determination of earthquake energy release and ML using TERRAscope. *Bull. Seismol. Soc. Am.* **83**, 330–346 (1993).
58. M. Kikuchi, Y. Yamanaka, K. Koketsu, Source process of the long-period seismic pulses associated with the 2000 eruption of Miyakejima volcano and its implications. *J. Geograph. Tokyo* **110**, 204–216 (2001). doi: 10.5026/jgeography.110.2_204


## ACKNOWLEDGMENTS

A large part of the field effort was supported with crisis response funding from the Icelandic government through the Civil Protection Department of the National Commissioner of the Icelandic Police. A considerable part of this work was made possible through the European Community's Seventh Framework Programme grant no. 308377 (Project FUTUREVOLC). Further support came from the EU Seventh Framework Marie Curie project NEMOH no. 289976 and the project CO2Volc ERC grant no. 279802, the Research Fund of the University of Iceland, the Irish Research Council, the Helmholtz Alliance on Remote Sensing and Earth System Dynamics (EDA), Bayerisches Geoinstitut through their DFG core facility for high pressure research, and UNAM/CIC Intercambio Académico. We thank the following key persons for help with instrumentation and data: B. H. Bergsson, V. S. Kjartansson, S. Steinthorsson, Th. Jónsson, P. Erlendsson, J. Soring, I. Jónsdóttir, G. Sverrisdóttir, and G. Sigurdsson; the South Iceland Lowland network monitoring team at the Icelandic Meteorological Office; the field observation team of the Institute of Earth Sciences in Holuhraun; and for collaboration on seismic instruments and data, we thank R. S. White and the Cambridge team and C. Bean and the University College Dublin team. The subsidence of Bárðarbunga was mapped by the surveying aircraft of Isavia (Icelandic Aviation Operation Services) and from satellite images acquired by Pléiades (Centre National d'Études Spatiales (CNES) CNES 2014, Distributed Airbus DS, subsidized access provided by the French Space Agency CNES's ISIS program) and SPOT 6 (Airbus DS 2014). Committee on Earth Observation Satellites (CEOS) and Geohazards Supersites and Natural Laboratories (GSNL) also provided access to satellite data. COSMO-SkyMed data were provided by the Italian Space Agency (ASI) through the Iceland Volcanoes Supersite agreement. The Icelandic Coast Guard provided valuable airplane and helicopter support for the field effort. Field operations at Bárðarbunga were supported by the Landsvirkjun power company, the Iceland Glaciological Society, and Vatnajökull National Park. Data are also available at http://futurevolc-www1.vedur.is.

## SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/353/6296/aaf8988/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S8
References (*59*–*96*)

16 April 2016; accepted 20 June 2016
10.1126/science.aaf8988

RESEARCH ARTICLES

OPTICS

# Shrinking light to allow forbidden transitions on the atomic scale

Nicholas Rivera,[1]*† Ido Kaminer,[1]* Bo Zhen,[2] John D. Joannopoulos,[1] Marin Soljačić[1]

**The diversity of light-matter interactions accessible to a system is limited by the small size of an atom relative to the wavelength of the light it emits, as well as by the small value of the fine-structure constant. We developed a general theory of light-matter interactions with two-dimensional systems supporting plasmons. These plasmons effectively make the fine-structure constant larger and bridge the size gap between atom and light. This theory reveals that conventionally forbidden light-matter interactions—such as extremely high-order multipolar transitions, two-plasmon spontaneous emission, and singlet-triplet phosphorescence processes—can occur on very short time scales comparable to those of conventionally fast transitions. Our findings may lead to new platforms for spectroscopy, sensing, and broadband light generation, a potential testing ground for quantum electrodynamics (QED) in the ultrastrong coupling regime, and the ability to take advantage of the full electronic spectrum of an emitter.**

A fundamental process in the field of light-matter interaction is spontaneous emission, in which an excited electron in an atom lowers its energy by emitting light (*1*–*3*). Spontaneous emission is responsible for the characteristic emission spectrum of an emitter. In principle, an excited electron can fall into any unoccupied lower energy level via this process. In practice, the majority of radiative decay channels are too slow to be accessible, rendering most of the spectrum invisible and inaccessible. The wealth of forbidden light-matter interaction processes can be appreciated by considering three very general classes of forbidden transitions: multipolar processes, spin-flip processes, and multiquanta emission processes.

Multipolar transitions are those in which the orbital angular momentum (OAM) quantum number of the electron changes by more than one unit. The transitions are slow because the wavelength of emitted light (around $10^3$ to $10^5$ Å) is typically far larger than the size of the atomic or molecular orbitals participating in the transition (around 1 to 10 Å). As a result of this difference in length scales, the rates of electronic transitions as a function of angular momentum vary over many orders of magnitude: The difference in rates between changing OAM by 1 and changing OAM by 5 can be more than 20 orders of magnitude, making high OAM transitions invisible in the absorption and emission spectrum of an emitter. The ability to access these transitions would allow a multiplex and broadband spectroscopy platform, in which much more of the electronic energy level structure of an emitter can be probed using light.

A spin-flip process is one in which the spin of the electron flips as a result of the emission. Common spin-flip processes include those in which a single electron flips its spin and those in which a pair interconverts between a singlet and triplet state. Because these processes are very slow, spin-flip processes typically occur through phosphorescence, which is fast only when the electrons experience large spin-orbit coupling (*4*, *5*). In systems where spin-orbit coupling is weak, such as in atoms or hydrocarbon molecules, the radiative quantum yield of phosphorescence is low because the singlet-triplet transition is much slower than nonradiative recombination processes. This is undesirable for applications such as organic-based light-emitting diodes where high radiative quantum efficiency is desired.

Multiquanta emission processes are those in which an electron lowers its energy by emitting two or more quanta of light. This is a higher-order process in QED. Although it is not forbidden by dipole selection rules, it is typically so slow that we group it with the other two processes above. Because of the small size of the atom relative to the light it emits, and the small value of the fine-structure constant, emission of even two quanta tends to be slower than the emission of a single quantum of light by 8 to 10 orders of magnitude. This makes two-quanta spontaneous emission processes notoriously difficult to observe. Two-photon spontaneous emission was predicted in the early 1930s (*6*) but was not directly observed until 1996, in hydrogen (*7*). Going beyond atomic physics, two-quanta spontaneous emission processes are also of interest in semiconductors, where they were first observed in 2008 (*8*, *9*). Enabling two-quanta spontaneous emission processes at a fast rate would enable generation of entangled light, which is of great interest for implementing quantum protocols (*10*). It would also enable new broadband absorbers and novel broadband light sources using atomic and molecular emitters.

Despite the motivation to access all of these transitions, it is in practice very difficult. Conservation of energy and angular momentum dictates that in order to access the full spectrum of an emitter, it must couple to light, which, over a broad range of frequencies in the infrared-visible range, has wavelengths on the order of 1 to 5 nm (once inside an optical material). The wavelength of light and its frequency are related by the phase velocity. Thus, to access forbidden transitions, phase velocities of light on the order of $0.001c$ are required. Recent theoretical and experimental work has shown that systems such as graphene, silver monolayers, and beryllium surfaces can support precisely such strongly confined light in the form of plasmons. These two-dimensional (2D) systems have the unique ability to squeeze the wavelength of light by more than two orders of magnitude, which allows us to recast the main assumptions for light-matter interactions.

We report that through the use of 2D plasmons (*11*–*24*), which shrink the wavelength of light by more than two orders of magnitude, we can overcome all of the mentioned conventional limitations of light-matter interactions. In particular, we show by means of a general theory of atom-plasmon interactions that the rates of high-order multipolar transitions, two-plasmon spontaneous emission, and spin-flip transition rates become enhanced to the point that they are comparable to dipole transition rates and thus become straightforwardly accessible. For achievable plasmon confinements in graphene and other 2D plasmon–supporting materials, all of the previously forbidden transitions listed above can now have lifetimes within the range of 1 μs to 1 ns; these previously forbidden transitions can be made to be quite efficient, even when competing with conventionally allowed transitions. We analyze the effect of nonradiative quenching extensively and find that in many cases of interest, the radiation into plasmons is dominant and thus could be harvested as far-field light through suitable outcoupling techniques.

## 2D plasmons and light-matter interaction: Model and assumptions

Plasmons are understood in the framework of classical electrodynamics as the coherent propagation of surface charge associated with tightly confined electromagnetic fields. The dispersion relation of any plasmon can be most simply expressed as

$$\omega = v_p q \equiv \frac{cq}{\eta(q)} \qquad (1)$$

where $q$ is the plasmon wave vector, $v_p$ is the phase velocity, and $\eta$ is the confinement factor, which is equal to both $c/v_p$ and $\lambda_0/\lambda_{pl}$. The fact peculiar to 2D plasmons is that their wavelength $\lambda_{pl}$, which is determined by the dispersion relation, can be shorter than the wavelength of a free-space

[1]Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: nrivera@mit.edu

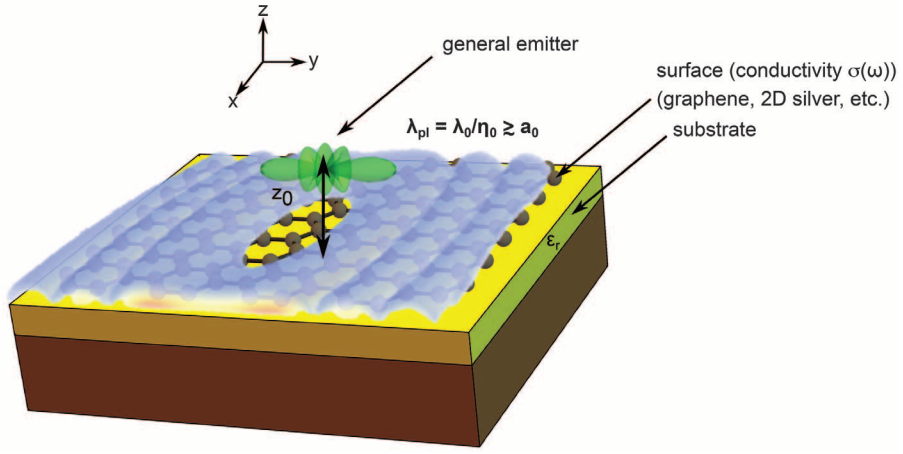



**Fig. 1. Two-dimensional plasmons coupled to an emitter.** A schematic of an emitter (not necessarily dipolar) above a 2D material of conductivity σ(ω) supporting plasmons with wavelength $\lambda_{pl}$, which is far shorter than the photon wavelength $\lambda_0$ (by a factor of $\eta_0$) and approaches the atomic size $a_0$.

photon $\lambda_0$ at the same frequency by a factor of 100 to 400.

One 2D plasmonic platform of recent interest is graphene, in which the confinement factor has been predicted to be as high as 300 (*15*); values in excess of 200 have been indirectly observed below the intraband regime (*23*), values of 150 have been directly observed with low losses (*21*), and values of 240 have been observed in the interband regime (*13*). Yet other 2D plasmonic platforms exist; what is less well known is that even more highly confined plasmons can exist in other 2D electron systems such as metallic monolayers. In other 2D plasmons such as monolayer silver, confinement factors as high as 300 have been observed, corresponding to plasmon wavelengths of ~5 nm at photon wavelengths of 1.5 μm (*11*). On the surface of beryllium, acoustic plasmons have been observed at visible frequencies (~560 nm) with plasmon wavelengths of merely 1.5 nm (*12*), corresponding to confinement factors near 370. In our study, we look at the emission of 2D plasmons by hydrogen-like and helium-like atomic emitters in the vicinity of a 2D plasmonic system given by the minimally coupled Hamiltonian

$$H = H_a + H_{em} + H_{int}$$

$$H_a = \left(\sum_i \frac{\mathbf{p}_i^2}{2m_e} - \frac{e^2}{4\pi\varepsilon_0 r_i}\right) + H_{e\text{-}e} + H_{SO}$$

$$H_{em} = \sum_{j=x,y,z} \int d\mathbf{r} \int d\omega\, \hbar\omega \left[f_j^\dagger(\mathbf{r},\omega) f_j(\mathbf{r},\omega) + \frac{1}{2}\right]$$

$$H_{int} = \sum_i \frac{e}{2m_e}[\mathbf{p}_i \cdot \mathbf{A}(\mathbf{r}_i) + \mathbf{A}(\mathbf{r}_i) \cdot \mathbf{p}_i] + \frac{e^2}{2m_e}\mathbf{A}^2(\mathbf{r}_i) + \frac{e\hbar}{2m_e}\sigma_i \cdot \mathbf{B}(\mathbf{r}_i) \quad (2)$$

where $e$ is the electronic charge, $m_e$ is the electron mass, $\hbar$ is the Planck constant divided by $2\pi$, $\mathbf{A}$ and $\mathbf{B}$ are the vector potential and magnetic field, $\mathbf{r}_i$ denotes the position of the $i$th electron, $\sigma_i$ denotes the spin of the $i$th electron, $H_{SO}$ is the spin-orbit coupling, and $H_{e\text{-}e}$ is the electron-electron interaction. The $f_j^\dagger(\mathbf{r},\omega)$ and $f_j(\mathbf{r},\omega)$ are operators that respectively create and annihilate elementary excitations of the electromagnetic field in dissipative media. These excitations can be thought of as oscillating dipoles at position $\mathbf{r}$ and frequency ω, oriented along direction $j$. The features of the plasmon and emitter relevant to our calculations are shown schematically in Fig. 1. This Hamiltonian, with field operators that appropriately describe the (potentially very high) losses in these confined plasmons, is sufficient to capture the fundamentally quantum nature of plasmons, which has been confirmed by experiments (*25*). This Hamiltonian describes both radiative decay and nonradiative decay mediated by interactions between the electromagnetic field and the orbital and spin degrees of freedom of the electron in an atom. It therefore describes intercombination transitions, multiplasmon emission, and other higher-order processes in the perturbation theory. In particular, this formalism makes predictions that (among others) give the correct Purcell factors for one-photon spontaneous emission of dipole emitters (*26*, *27*), describe experimentally observed loss-induced quenching phenomena (*20*, *28*), and describe experimentally observed two-plasmon phenomena as in (*9*).

We examined all of these interaction processes in order to provide a broad picture of atom-light interactions in plasmonics (*29*). We selected hydrogen- and helium-like emitters for definiteness, but the basic physics behind our results can readily be extended to any other atomic and molecular system. Although we explicitly consider only spontaneous emission, our results hold equally well for the absorption of excited 2D plasmons. Our approach can be applied to plasmons in 2D conductors with any general dispersion relation (for which the Drude model is a representative example that we consider). Our model can be generalized to consider nonlocal effects (*30*) by modifying the Green function formalism, which is done by replacing the local conductivity σ(ω) by the nonlocal conductivity σ(*q*, ω) everywhere. For the most part, our claims can be chosen to apply within the local-response regime of graphene, for easily achievable values of the Fermi energy $E_F$ of graphene. We explicitly consider the nonlocal response of graphene [via the zero-temperature random phase approximation (RPA) (*15*, *27*)] when we discuss two-plasmon spontaneous emission.

## Multipolar transitions

The first class of forbidden transitions that we analyze consists of multipolar transitions in which the OAM of the electron changes by more than 1. It is well known that in conventional photonics and even plasmonics, electric dipole transitions are the fastest type of transitions and are thus appropriately the main object of study in the field of light-matter interactions. Electric quadrupole and magnetic dipole transitions, which are the second fastest types of transitions, are fairly slow but can be observed through conventional spectroscopic approaches (*31*) or in mesoscopic emitters such as quantum dots. Substantial theoretical and experimental efforts have gone toward speeding up these two processes (*32*–*39*). On the other hand, electric octupole, magnetic quadrupole, and higher-order multipolar transitions are far too slow to observe for small emitters such as atoms and molecules. Because of this, appreciable theoretical and experimental efforts have not been directed at speeding up such processes. Nevertheless, the spectrum of an atom associated with high-lying energy levels consists of precisely these types of transitions. Because of the high confinements achievable in 2D plasmons, transitions whose lifetimes in free space approach the age of the universe can be brought down to lifetimes of hundreds of nanoseconds, corresponding to a rate enhancement (defined as the Purcell factor) of nearly $10^{24}$, as we demonstrate below. Such lifetimes are of the same order of magnitude as dipole transitions in free space and therefore should be straightforwardly accessible through absorption spectroscopy.

We calculate the rates of transitions where the electron OAM changes by $n$ (an E$n$ transition) (*29*). The decay rate $\Gamma_n$ into plasmons scales with $\eta_0$ as

$$\Gamma_n \sim \eta_0^{3+2(n-1)} \exp(-2\eta_0 k z_0) \quad (3)$$

where $k = 2\pi/\lambda_0$, and $\eta_0$ is the confinement factor at the transition frequency. Equation 3 makes it very clear that higher electric multipole transitions are enhanced by successively larger amounts. For achievable confinement factors of 100, when $n$ increases by 1, the enhancement increases by a factor of roughly 10,000. Therefore, one should expect that if the dipole transitions ($n$ = 1) are enhanced by $10^6$, then E2, E3, E4, and E5 transitions should be enhanced by factors of $10^{10}$, $10^{14}$, $10^{18}$, and $10^{22}$, respectively. One feature of Eq. 3 is that for high-order multipolar transitions, the decay rates have an extremely sharp dependence on the confinement. For example, with E5 transitions, where the rates of surface plasmon launching scale as $\eta_0^{11}$, a

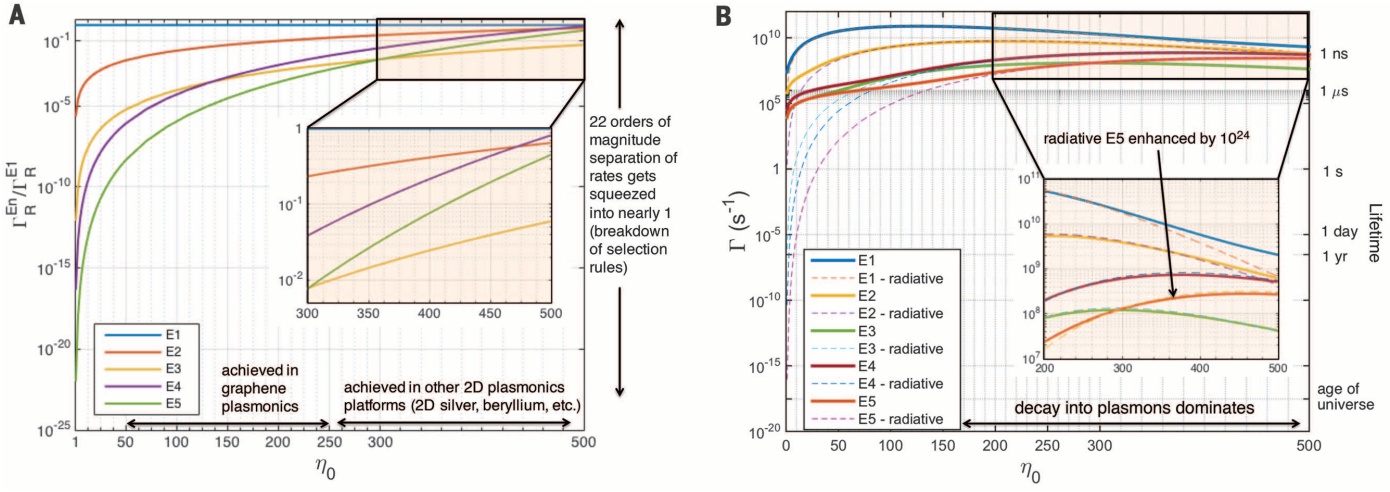



**Fig. 2. Convergence of the multipoles.** (**A**) Rates of radiation into surface plasmons for various multipolar transitions in hydrogen. The transition series considered here is 6{p, d, f, g, h} → 4s, which are E1 to E5 transitions, respectively. The free-space wavelengths of the transitions are all 2.6 μm. The emitter is situated 5 nm above the surface of graphene. (**B**) Total rates (radiative + nonradiative) of decay for the same transition series in hydrogen (for a fixed quality factor of 5). Dashed lines show radiative rates, which agree with the total rates at high confinement.

change in the confinement factor by a factor of 2 can change the decay rates by more than three orders of magnitude.

The expression above can be intuitively reasoned as follows: The factor $\eta_0^3 \exp(-2\eta_0 kz_0)$ is proportional to the local photonic density of states (LDOS), reflecting the well-known fact that plasmon confinement can lead to very fast dipole transitions (*26*, *27*, *40*). The additional factors of $\eta_0$ that depend on the multipolar order of the transition are field-gradient enhancement factors. In particular, if we expand the plasmonic plane wave as

$$\exp(i\mathbf{q}\cdot\rho - qz) = 1 + (i\mathbf{q}\cdot\rho - qz) + \frac{1}{2!}(i\mathbf{q}\cdot\rho - qz)^2 + \ldots \quad (4)$$

we see that each term, which corresponds to some number of gradients of the electromagnetic field, can be mapped to particular multipolar transitions for wavelengths not too close to the atomic radius. Because $q = \eta_0(\omega/c)$, it is clear that each multipolar order is enhanced (relative to the dipole order) by some even power of $\eta_0$ and that this $\eta_0$ is proportional to the gradient of the electromagnetic field.

Such an intuition is confirmed in Fig. 2, where we plot the exact transition rates, with and without plasmonic losses, for the series of transitions 6{p, d, f, g, h} → 4s in the hydrogen atom. The free-space wavelength of the transition is 2.6 μm. In our simulations, the emitter is kept 5 nm away from the surface. We do not include the 6s → 4s rate because, for the parameters under consideration, it is extremely small when the atom is well separated from the surface because of the approximate transversality of the electromagnetic field ($\nabla \cdot \mathbf{E} = 0$) (*41*). The dominant decay mechanism for this process will be two-plasmon emission, which we compute below. In Fig. 2A, we plot the radiative rates (i.e., plasmon emission rates, denoted $\Gamma_R$) of transitions in the hydrogen 6 → 4 series relative to the radiative rate of the dipole transition (at the same energy) as a function of confinement. This relative rate is independent of atom-surface separation, as can be seen from Eq. 3. To give a particular example, conventionally the E5 transition is separated from the E1 transition by ~22 orders of magnitude. The free-space rate of the E5 transition is around $10^{-16}$ $s^{-1}$, corresponding to a lifetime that is two orders of magnitude less than the age of the universe. In contrast, for high confinement ($\eta_0 \approx 200$), the rate is ~$10^7$ $s^{-1}$, which is more than an order of magnitude faster than the free-space E1 transition. Moreover, at this confinement, the entire multipolar series is separated by only three orders of magnitude (in rates), as opposed to 22 orders of magnitude in free space. A confinement of 200 is easily achievable in graphene in its local-response regime. For example, by tuning $E_F$ to 0.4 eV, this 0.47-eV transition has a confinement of 200 when placed above a substrate with a relative permittivity of 4. Thus, it is possible to allow highly forbidden transitions in the local regime of graphene.

In addition to plotting confinement factors up to 200, we also plot confinement factors that are higher than have been observed in graphene below the interband regime. This is because such high confinement factors have been observed in plasmons in monolayer silver, $DySi_2$, graphene with high losses, and acoustic plasmons in beryllium (*11*–*14*). We plot up to very high confinement factors of 500, corresponding to plasmon wavelengths of 5 nm. At these low wavelengths, which have been observed in the aforementioned systems, four of the five transitions in the series lie within the same order of magnitude, corresponding to a complete breakdown of the conventional selection rules. Another reason to plot for high confinement factors (even for graphene) is that it has recently been suggested via a modified RPA scheme that confinements above 300 may be possible with relatively low losses (*42*). Technically, in these highly confined plasmons with linear dispersion, the quality factors $Q$ have been observed to be quite low (of order unity). Figure 2B assumes a fixed $Q$ of 5 throughout different confinement factors. If $Q = 1$, rates are, of course, quantitatively modified but remain within the same order of magnitude as in Fig. 2B, preserving our conclusion that these highly confined plasmons can break the selection rules completely. Moreover, it tells us that even despite very high losses of these plasmons, they can still completely reshape the dynamics of excited electrons in emitters. Therefore, these highly lossy plasmons have clear experimental consequences, even if they cannot be coupled out into free space. We discuss these consequences below.

In addition to radiation into plasmons, an excited emitter can be nonradiatively quenched by a conductor. Nonradiative quenching channels can include, but are not limited to, phonons, particle-hole excitations, and impurities. These are incorporated in our macroscopic QED calculations through the real part of the surface conductivity. Strictly speaking, because macroscopic QED sees only the total conductivity, it considers only the total quenching (by which we mean total nonradiative decay). Therefore, we cannot isolate the quenching sources, but we may still compare the total quenching to the total plasmon emission. In Fig. 2B, we plot the total rates of the transitions considered in Fig. 2A [i.e., radiative + nonradiative (solid lines) in addition to the radiative decay rates (dashed lines)]. For low confinement, the nonradiative decay is completely dominant. This reflects the well-known fact that nonradiative energy transfer strongly depends on the ratio $1/(\eta_0 kz_0)$ (*29*). A similar result of quenching was pointed out in (*43*). What is more valuable for optical applications is the situation at high confinement. The threshold value of confinement in order to be considered "high confinement" depends on the multipolar order in question. For dipole transitions, high confinement can mean $\eta_0 > 20$. For E5 transitions, it can mean $\eta_0 > 150$.

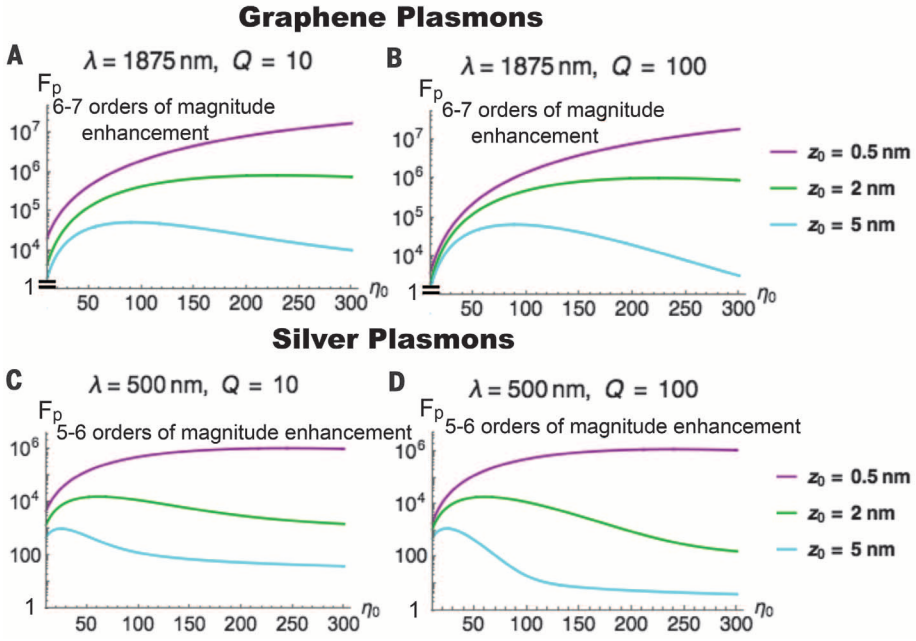



**Fig. 3. Enabling singlet-to-triplet transitions.** (**A** to **D**) Purcell factors for a dipolar singlet-triplet transition as a function of confinement factor $\eta_0$ for plasmons in graphene [(A) and (B)] and plasmons in 2D silver [(C) and (D)] for different atom-plane separations $z_0$ and quality factors $Q$. These results also apply to other kinds of intercombination transitions, such as those in which the spin of an atomic emitter is flipped because of radiation.

In the high-confinement regime, the radiation into plasmons is dominant and can be harvested as far-field light through suitable outcoupling techniques such as gratings and nanoantennas (*39*, *44*–*47*).

### Spin-flip transitions and the singlet-to-triplet transition

We next analyze those transitions for which the initial state does not directly connect to the final state by the emission of light, but rather couples to a virtual state whose symmetry is compatible with that of the final state. Take as an example a radiative transition between a spin-singlet state $S$ and a spin-triplet state $T$. The dominant process for systems with large spin-orbit coupling is a second-order process in which an initial triplet state connects to a virtual singlet state $S_n$ through the spin-orbit coupling, and then the virtual state connects to the final singlet state and emits light. The decay rate is then given by the second-order perturbation theory expression

$$\Gamma(T \rightarrow S) = \frac{2\pi}{\hbar^2}\sum_{\mathbf{q}}\left|\sum_n \frac{\langle S,\mathbf{q}|H_{\mathrm{em}}|S_n,0\rangle\langle S_n,0|H_{\mathrm{SO}}|T,0\rangle}{E_T - E_{S_n}}\right|^2 \times \delta(\omega_{\mathbf{q}} - \omega_0) \quad (5)$$

where $\omega_0$ is the transition frequency and $\mathbf{q}$ is the wave vector of an emitted plasmon.

Use of the Fermi Golden Rule to calculate the decay rate generally involves summation over all possible intermediate states. However, in many cases, there is an intermediate singlet state that is nearly degenerate with the triplet state, which then gives the dominant contribution to the transition, allowing us to ignore the existence of other singlet states. The advantage of this approximation is that the spin-orbit and electromagnetic enhancements decouple. Our approach allows the calculation of the enhancement factor without knowing anything about the spin-orbit coupling, or even about the structure of the atom or molecule, and thus our results can be applied to many atomic and molecular systems. The Purcell factor, defined as the emission rate into plasmons divided by the emission rate into free-space photons, in the absence of plasmon losses, is proportional to

$$F_{\mathrm{p}}(T \rightarrow S) \sim \eta_0^3 \exp(-2\eta_0 k z_0) \quad (6)$$

where $\eta_0$ is the confinement factor at the transition frequency $\omega_0$, and $z_0$ is the separation between the atomic nucleus and the surface. For the same reason as discussed previously, the mechanism of this enhancement is purely LDOS-based. This simple formula is only weakly modified by plasmonic losses (for reasons discussed above), which were taken into account in our numerical calculations (*29*). These high enhancement factors are in agreement with the $\eta_0^3$ decay law of pure electric dipole transitions (*26*, *27*).

In Fig. 3, we illustrate the Purcell factors achievable for an emitter on top of graphene and monolayer silver as a function of plasmon confinement. We consider these transitions in a helium-like atom, where the singlet and triplet states (split by electron repulsion) are understood as coming from pairs of one-electron states (*29*). We show the Purcell factors at distances of 0.5, 2, and 5 nm away for two different plasmon quality factors ($Q$ = 10 and $Q$ = 100). Figure 3 demonstrates how 2D plasmons can enhance singlet-to-triplet transition by up to seven orders of magnitude for realistic and experimentally readily available parameters. This enhancement is separate from the fact that additional enhancements can come from increasing spin-orbit coupling (*5*). This allows one to combine approaches involving high spin-orbit coupling and approaches involving plasmonics, thereby allowing attainment of the highest possible enhancements. Perhaps of more interest, it allows one to reconsider the use of emitters with weak spin-orbit coupling. For example, by enhancing the triplet decay by four to six orders of magnitude, one can take a triplet state of a hydrocarbon with a typical phosphorescence time of 0.1 to 0.001 s and make it competitive with, or even faster than, phosphorescence of organometallic triplet states with high spin-orbit coupling.

Another method of enhancing spin-flip and singlet-triplet transitions involves a direct emission into a plasmon through the $\mathbf{S}\cdot\mathbf{B}$ term of the interaction Hamiltonian. Such a direct spin-flip (or singlet-triplet) transition involves an M$n$ transition, for which the decay rate into plasmons is slower than its electric counterpart by the ratio $\hbar\omega_0/m_e c^2 \approx 10^{-5}$ to $10^{-6}$ and is thus much slower than the intercombination mechanism for spin-orbit couplings of interest. We demonstrate this claim by deriving rates for M$n$ transitions (*29*).

### Two-plasmon spontaneous emission

The emission of two excitations of a field (photons or plasmons) is a second-order effect in perturbation theory. If the dimensionless coupling constant $g$ of atomic QED is small, then the two-quanta emission will be very slow relative to the emission of a single field excitation. A good estimate for the coupling constant is given by $g^2 = \Gamma/\omega_0$. When the value of $g^2$ approaches 1, we expect the Fermi Golden Rule (and more generally, perturbation theory) to fail, because the width of the resulting atomic resonance is comparable to the frequency scale of variation of the continuum matrix elements (*48*). Taking this as our dimensionless coupling constant, we see that it scales as

$$g_n \sim \sqrt{\alpha(ka_0)^{2n}\eta_0^{3+2(n-1)}\exp(-2\eta_0 k z_0)} \quad (7)$$

in atomic and molecular systems, where $a_0$ is the characteristic emitter size (for example, the Bohr radius).

For dipole transitions, this coupling constant suggests that the emission rate of two plasmons scales as $g^4 \sim \eta_0^6$. Our rigorous derivation shows that this is indeed the case (*29*) for both Drude-like and linear plasmon dispersions. Similar conclusions hold for other dispersion relations. We arrive at the following analytical expression for the enhancement of the differential decay rate

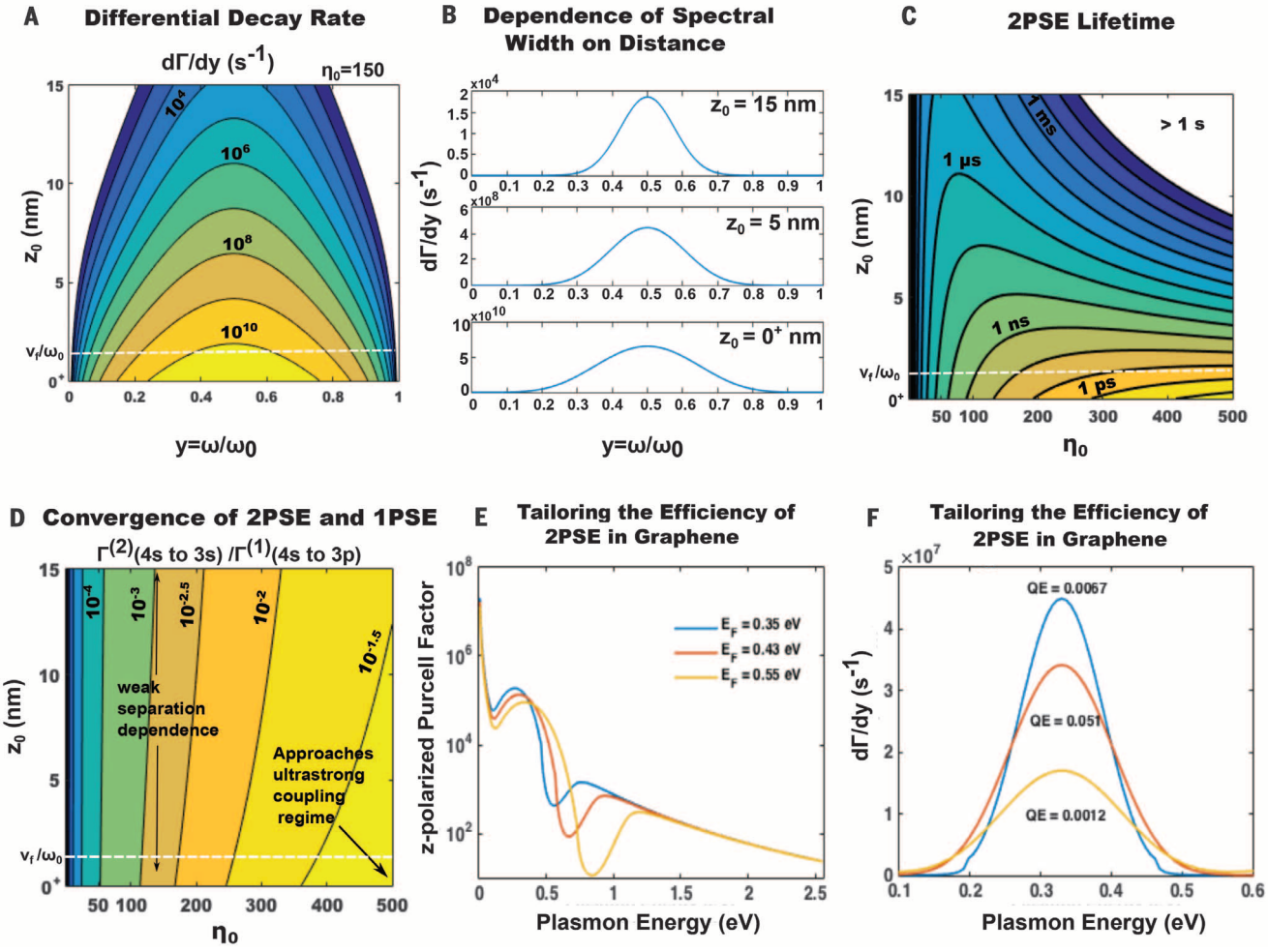



**Fig. 4. Enabling two-plasmon radiative energy transfer.** (**A**) Emission spectrum of two-plasmon spontaneous emission (2PSE) as a function of frequency and atom-surface separation for the hydrogen 4s → 3s transition above graphene. $\eta_0$ is chosen to be 150. (**B**) Line cuts of (A) for atom-surface separations of 0, 5, and 15 nm. (**C**) Total decay rate (in $s^{-1}$) for this transition as a function of confinement and atom-surface separation. (**D**) Comparison of two-plasmon emission rate to the emission rate for a single plasmon 4s → 3p transition as a function of confinement and separation. (**E**) Purcell spectra for a *z*-polarized dipole 10 nm from graphene as a function of transition energy for different values of the Fermi energy of graphene. (**F**) Two-plasmon spontaneous emission spectra as a function of the energy of one of the emitted plasmons for the same values of Fermi energy as in (E). The plots show the transition lifetime and the quantum efficiency of the transition, assuming that the competing transitions (4s → 3p and 4s → 2p) are dipolar.

(in the lossless limit) for any two-plasmon $S \rightarrow S$ transition in a hydrogenic atom:

$$\frac{d\Gamma/dy|_{2pl}}{d\Gamma/dy|_{2ph}} = \frac{d\Gamma/d\omega|_{2pl}}{d\Gamma/d\omega|_{2ph}} = 72\pi^2[\eta_0^3 \exp(-2\eta_0 kz_0)]^2 \times (y - y^2)^3 \exp[8\eta_0 kz_0 y(1 - y)] = \frac{1}{2}F_p(y)F_p(1 - y) \quad (8)$$

where $y = \omega/\omega_0$, $\eta_0$ is the confinement factor at the spectral peak of the emission, and $F_p$ is the Purcell factor for one-plasmon spontaneous emission for a dipole polarized perpendicular to the surface of the 2D conductor. This peak occurs for $\omega = \frac{1}{2}\omega_0$ or, equivalently, $y = \frac{1}{2}$. $d\Gamma/d\omega(\omega')$ corresponds to the rate of emission of two plasmons per unit frequency, in which one plasmon is at frequency $\omega'$ and the other is at frequency $\omega_0 - \omega'$. A quick estimate using Eq. 8 reveals that at confinement factors of 200, the left side of Eq. 8, defined as the spectral enhancement, tends to $10^{15}$ as $z_0 \rightarrow 0$. The $\frac{1}{2}F_p(y)F_p(1 - y)$ term of Eq. 8 reveals the mechanism of two-plasmon spontaneous emission enhancement: LDOS enhancement over a sufficiently broad frequency spectrum such that enhancement at complementary frequencies, $\omega$ and $\omega_0 - \omega$, is very high. For Fig. 4, A to D, we consider two-plasmon spontaneous emission without considering losses. We discuss the effect of plasmon losses below.

In Fig. 4A, we plot the distribution of emitted plasmons as a function of normalized frequency, *y*. The spectral distribution narrows for increasing atom-surface separation $z_0$ (Fig. 4B). This arises from the *y* dependence of the exponential in Eq. 8. Therefore, through a precise control of the position of an emitter above a surface supporting plasmons, it is possible to tune not only the rate of emission, but also the spectrum of emission.

In Fig. 4C, we plot an estimate of the total decay rate for the transition 4s → 3s, which is computed by summing over 10 virtual discrete states (*29*). The rate converges sufficiently for 10 virtual states. Even for modest confinements of around 100, and at separations of 5 nm, the two-plasmon spontaneous emission rate approaches 0.1 $ns^{-1}$, in stark contrast to the typical rate of roughly 1 $min^{-1}$ in free space. For confinements beyond 200, it is possible to obtain plasmon emission rates exceeding 0.1 $ps^{-1}$. We compare the rate of this transition with that of the 4s → 3p single-plasmon dipole transition in Fig. 4D; in the region of extreme confinement (200 to 500), the two-plasmon emission is only one to two orders of magnitude slower. At these extreme confinements—such that the two-plasmon emission is slower than the

one-plasmon emission by two orders of magnitude—the dimensionless coupling constant $g$ is on the order of 0.1, signaling the onset of the ultrastrong coupling regime.

We find that for plasmons described by the Drude model, the ratio of the rates has a very weak distance dependence. This arises from the fact that at greater atom-surface separations, the spectrum of emitted plasmons gets narrower. Therefore, most of the emitted plasmons have frequencies near $y = ½$. When this happens, the distance dependence of the two-plasmon decay rate approaches $\exp(-2\eta_0 kz_0)$, which is the same distance dependence as that of the one-plasmon decay rate. This is a feature of the $\omega \sim \sqrt{q}$ dispersion of plasmons, which is well described by the Drude model. For plasmons with linear dispersion (such as acoustic plasmons), the distance dependence of the two-plasmon differential emission rate is simply $\exp(-4\eta_0 kz_0)$ (i.e., no $y$ dependence, in contrast to that of Eq. 8). The results summarized in Fig. 4, A to D, establish that two-plasmon spontaneous emission is very fast in the vicinity of 2D plasmon–supporting materials. Indeed, it should already be experimentally achievable to observe lifetimes for two-plasmon processes on the order of 1 ns, which is of similar order of magnitude to that of fast dipole transitions in free space. We note that the effect of losses is considered in our study (*29*). In brief, for this particular transition, second-order quenching effects are only of extreme relevance at very short atom-surface separations of below 1 nm (see Fig. 4, E and F, and fig. S4 to see how the results in Fig. 4, A to D, are modified by losses). However, our consideration of losses opens up the possibility of observing and potentially using a new decay mechanism for emitters near a surface: nonradiative decay through two (entangled) lossy excitations (i.e., double quenching).

Next, we show that it is even possible to exceed the efficiencies for two-plasmon emission suggested by Fig. 4D, allowing for the possibility of efficient sources of entangled plasmons. It is known both theoretically (*27*) and experimentally (*20*) that graphene features a spectral region (roughly) between $E_F$ and $2E_F$ in which the Purcell factor drops by several orders of magnitude. We call this region the "dip region." Consider a hydrogenic electron in the 4s state. The 4s can decay into the 2p and 3p states by first-order emission. It can also potentially decay into 3s via two-plasmon spontaneous emission. All other processes are negligible. Suppose that the energy difference between 4s and 3s (0.66 eV) lies within the spectral dip region. Further suppose that highly confined plasmons exist at 0.33 eV. Then, the relative enhancement between second order and first order emission processes will be much higher than suggested by Fig. 4D, which assumed no such dip. In Fig. 4E, we show the Purcell factor for first-order (dipolar) emission for a $z$-polarized dipole at atom-surface separation of 10 nm for graphene Fermi energies of 0.35, 0.43, and 0.55 eV. Our calculations take the spatial nonlocality of graphene's permittivity into account via the RPA, as is done in (*15*, *27*). In Fig. 4F, we show the spectrum of plasmon emission (which is integrated to get the overall transition rates) and quantum efficiencies for two-plasmon spontaneous emission at these different atom-surface separations (*49*). The quantum efficiency here is well approximated by

$$\mathrm{QE} = \frac{\Gamma(4s \rightarrow 3s)}{\Gamma(4s \rightarrow 3s) + \Gamma(4s \rightarrow 3p) + \Gamma(4s \rightarrow 2p)} \quad (9)$$

We find that the quantum efficiency of two-plasmon emission is an extremely strong function of Fermi energy, jumping by an order of magnitude as $E_F$ is varied from 0.35 eV to 0.43 eV and as $E_F$ is varied from 0.43 eV to 0.55 eV. At $E_F$ = 0.43 eV, the quantum efficiency of two-plasmon emission is 5%, which is not only quite high for a second-order process, but is also far larger than what is suggested by Fig. 4D, and it is enabled as a result of the spectral dip in graphene. Therefore, using low-loss plasmons such as those observed recently in graphene (*21*) and the results presented here, one may be able to generate and couple out entangled plasmons at a fast rate and at a relatively high efficiency. These findings suggest very clearly that it should be possible to use systems with high Purcell factors over a narrow spectral region to enhance the relative efficiency of multipolar and spin-flip transitions as well (although not in the highly degenerate hydrogen atom). Other systems with high Purcell factors over a narrow spectral region through which our findings may be realized include nanometer-scale plasmonic resonators with high quality factors and few-nanometer-thick metallic thin films.

## Summary and outlook

We have shown that 2D plasmonics, with its unprecedented level of confinement, presents a unique opportunity to access radiative transitions in atomic-scale emitters that were considered inaccessible. Multiplasmon processes, spin-flip radiation, and very high-order multipolar processes have been shown to be competitive with the fastest atomic transitions in free space, rendering them all accessible. These findings (Table 1) pave the way to observing and taking advantage of new light-matter interactions. In particular, our findings may force the reconsideration of many atomic, molecular, and solid-state emitters whose transitions are normally too weak and inefficient to be used, thus bringing the full variety of the periodic table to optical applications. The physics responsible for the effects reported here may also be present in optical antennas whose dimensions are on the order of 1 nm (*50*). However, even if they are, an advantage of planar 2D plasmonics for accessing forbidden transitions is that the effect does not depend on the position of the emitter in the plane parallel to the 2D conductor. In cavity-based emission enhancement platforms, especially those in which fields are confined in a few-nanometer gap, the (atomic-scale) emitter must be positioned very precisely in order to observe the enhancement.

In many cases, we found that quenching was a weak correction to the decay of the emitters and that the primary decay was into propagating (albeit lossy) plasmons, which could be coupled out into far-field photons. Thus, the presence of forbidden transitions can be translated into far-field photons at frequencies not in the conventional emission spectrum of the emitter. Thus, one could experimentally observe forbidden transitions (and infer their rates) indirectly through these unconventional far-field photons. Note that even if the plasmons could not be outcoupled—for example, because of debilitatingly high losses—there will often be clear experimental signatures of these transitions. In particular, because some of the transitions will primarily occur via emission of a far-field photon, the presence of forbidden decay pathways will modify the far-field spectrum of an emitter, and a quantitative knowledge of the modified far-field emission spectrum of the emitter will allow one to infer the presence of forbidden transitions and the type of forbidden transition taking place. In fig. S5, we show

**Table 1. A summary of derived scalings of rates and rate enhancements for emitters on top of lossless 2D plasmons at zero displacement from the plasmon-supporting surface.** The rates for an emitter at finite distance from the plasmon are suppressed by factors of $\exp(-2\eta_0 kz_0)$, which is of order unity for an emitter within a reduced plasmon wavelength of the surface. Note that these scalings are still a good approximation even in the presence of high losses, provided that the confinement is large (Fig. 2). 2PSE, two-plasmon spontaneous emission.

| Transition* | Free space Γ | 2D plasmon Γ | η Enhancement |
|---|---|---|---|
| E1 | $\alpha(\hbar\omega/m_e c^2)$ | $\alpha(\hbar\omega/m_e c^2)\eta^3$ | 3 |
| E$n$ | $\alpha(\hbar\omega/m_e c^2)(ka_0)^{2(n-1)}$ | $\alpha(\hbar\omega/m_e c^2)(ka_0)^{2(n-1)}\eta^{3+2(n-1)}$ | $3 + 2(n-1)$ |
| Spin-flip E$k$† | $\alpha(\hbar\omega/m_e c^2)(ka_0)^{2(n-1)}$ | $\alpha(\hbar\omega/m_e c^2)(ka_0)^{2(n-1)}\eta^{3+2(n-1)}$ | $3 + 2(n-1)$ |
| 2PSE (dipole) | $\alpha^2(ka_0)^4$ | $\alpha^2(ka_0)^4\eta^6$ | 6 |
| M1‡ | $\alpha(\hbar\omega/m_e c^2)^2$ | $\alpha(\hbar\omega/m_e c^2)^2\eta$ | 1 |
| Mn‡ | $\alpha(\hbar\omega/m_e c^2)^2(ka_0)^{2(n-1)}$ | $\alpha(\hbar\omega/m_e c^2)^2(ka_0)^{2(n-1)}\eta^{1+2(n-1)}$ | $1 + 2(n-1)$ |

*Losses can change these significantly at low confinement and/or short separation.
†Not including the spin-orbit matrix element, which approximately does not contribute to the Purcell enhancement.
‡M1 and Mn are discussed in (*29*).

an energy-level diagram in an emitter conducive to observing forbidden transitions, even without outcoupling.

With regard to applications of this work to fundamental light-matter interactions, we believe that a number of fruitful extensions are within reach. Beyond the processes that we considered in this work, one can consider combinations of these processes, such as (i) multiplasmon transitions mediated by higher-order multipole virtual transitions, such as a two-plasmon emission with a total angular momentum change of 4 for the electron by way of intermediate quadrupole virtual transitions; (ii) third- and higher-order plasmon emission and absorption processes; or (iii) a second-order absorption process in which a plasmon and a far-field photon are absorbed, leading to large changes in energy and angular momentum of the electron due to the photon and plasmon, respectively.

All of these transitions should be enhanced using 2D plasmonics, but beyond the reach of conventional plasmonics and photonics without extremely high intensities of light. We stress that the results we have presented are not optimized. We performed our work with hydrogen purely as a proof-of-concept. For example, by finding an atom with a high octupole moment, it should be possible to make the octupole transition dominant over the dipole and other multipolar transitions, thereby paving the way for emitters with characteristic multipolarity different from dipolar. It should similarly be possible to optimize the rate of multiplasmon spontaneous emission relative to other transitions, leading to emitters highly capable of emitting entangled light.

Potential applications of this work include spectroscopy for inferring electronic transitions that cannot be determined with photons, sensors based on forbidden transitions, organic light sources arising from fast singlet-triplet transitions, fast entangled light generation, and fast generation of broadband light with tunable width in the near- and mid-infrared. Many of these applications could also be realized with the highly dissipative visible-frequency excitations in 2D conductors.

## REFERENCES AND NOTES


1. P. A. Dirac, *Proc. R. Soc. London Ser. A* **114**, 243–265 (1927).
2. M. Pelton, *Nat. Photonics* **9**, 427–435 (2015).
3. M. Tame *et al.*, *Nat. Phys.* **9**, 329–340 (2013).
4. R. Friend *et al.*, *Nature* **397**, 121–128 (1999).
5. A. Köhler, H. Bässler, *Mater. Sci. Eng. Rep.* **66**, 71–109 (2009).
6. M. Göppert, *Naturwissenschaften* **17**, 932 (1929).
7. C. L. Cesar *et al.*, *Phys. Rev. Lett.* **77**, 255–258 (1996).
8. A. Hayat, P. Ginzburg, M. Orenstein, *Nat. Photonics* **2**, 238–241 (2008).
9. A. Nevet *et al.*, *Nano Lett.* **10**, 1848–1852 (2010).
10. A. Hayat, A. Nevet, P. Ginzburg, M. Orenstein, *Semicond. Sci. Technol.* **26**, 083001 (2011).
11. T. Nagao, T. Hildebrandt, M. Henzler, S. Hasegawa, *Phys. Rev. Lett.* **86**, 5747–5750 (2001).
12. B. Diaconescu *et al.*, *Nature* **448**, 57–59 (2007).
13. Y. Liu, R. Willis, K. Emtsev, T. Seyller, *Phys. Rev. B* **78**, 201403 (2008).
14. E. Rugeramigabo, T. Nagao, H. Pfnür, *Phys. Rev. B* **78**, 155402 (2008).
15. M. Jablan, H. Buljan, M. Soljačić, *Phys. Rev. B* **80**, 245435 (2009).
16. S. J. Park, R. E. Palmer, *Phys. Rev. Lett.* **105**, 016801 (2010).
17. A. Grigorenko, M. Polini, K. Novoselov, *Nat. Photonics* **6**, 749–758 (2012).
18. Z. Fei *et al.*, *Nature* **487**, 82–85 (2012).
19. M. Jablan, M. Soljačić, H. Buljan, *Proc. IEEE* **101**, 1689–1704 (2013).
20. K. Tielrooij *et al.*, *Nat. Phys.* **11**, 281–287 (2015).
21. A. Woessner *et al.*, *Nat. Mater.* **14**, 421–425 (2015).
22. Q. Zhang *et al.*, *Sci. Rep.* **4** 6559 (2014).
23. Z. Fei *et al.*, *Nano Lett.* **11**, 4701–4705 (2011).
24. H. Yan *et al.*, *Nat. Photonics* **7**, 394–399 (2013).
25. E. Altewischer, M. P. van Exter, J. P. Woerdman, *Nature* **418**, 304–306 (2002).
26. A. Archambault, F. Marquier, J.-J. Greffet, C. Arnold, *Phys. Rev. B* **82**, 035411 (2010).
27. F. H. Koppens, D. E. Chang, F. J. García de Abajo, *Nano Lett.* **11**, 3370–3377 (2011).
28. L. Gaudreau *et al.*, *Nano Lett.* **13**, 2030–2035 (2013).
29. See supplementary materials on *Science* Online.
30. We expect nonlocal effects to kick in at atom-surface separations $z_0 \leq v_f/\omega_0$, where $v_f$ is the Fermi velocity and $\omega_0$ is the transition frequency.
31. A. Kramida *et al.*, Atomic Spectra Database (National Institute of Standards and Technology, Gaithersburg, MD, 2012).
32. J. R. Zurita-Sánchez, L. Novotny, *J. Opt. Soc. Am. B* **19**, 1355 (2002).
33. J. R. Zurita-Sánchez, L. Novotny, *J. Opt. Soc. Am. B* **19**, 2722 (2002).
34. R. Filter, S. Mühlig, T. Eichelkraut, C. Rockstuhl, F. Lederer, *Phys. Rev. B* **86**, 035404 (2012).
35. M. Takase *et al.*, *Nat. Photonics* **7**, 550–554 (2013).
36. V. Yannopapas, E. Paspalakis, *J. Mod. Opt.* **62**, 1435–1441 (2015).
37. M. L. Andersen, S. Stobbe, A. S. Sørensen, P. Lodahl, *Nat. Phys.* **7**, 215–218 (2011).
38. I. D. Rukhlenko *et al.*, *Opt. Express* **17**, 17570–17581 (2009).
39. P. K. Jain, D. Ghosh, R. Baer, E. Rabani, A. P. Alivisatos, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 8016–8019 (2012).
40. G. M. Akselrod *et al.*, *Nat. Photonics* **8**, 835–840 (2014).
41. Technically, $\nabla \cdot \mathbf{E} \neq 0$ only at the surface, but the contribution is small because the atomic wave functions are exponentially suppressed at the surface, so for parameters of interest, the direct s → s transition is much weaker than the two-plasmon process.
42. A. F. Page, F. Ballout, O. Hess, J. M. Hamm, *Phys. Rev. B* **91**, 075404 (2015).
43. A. Alabastri, X. Yang, A. Manjavacas, H. O. Everitt, P. Nordlander, *ACS Nano* **10**, 4835–4846 (2016).
44. G. Liu *et al.*, *ACS Nano* **6**, 6786–6792 (2012).
45. A. N. Abbas *et al.*, *ACS Nano* **8**, 1538–1546 (2014).
46. T. B. Hoang *et al.*, *Nat. Commun.* **6**, 7788 (2015).
47. For applications in which fast nonradiative decay is desired, we note that for considerably low wavelengths $1/kz_0 \approx 1$, the nonradiative decay is much faster than one would naively expect. This is because the nonradiative decay still couples to the multipole moment relevant to the atomic transition. Therefore, the strong enhancement of rates coming from matching the atomic size to the wavelength of light compensates for the suppression of rates coming from a higher $kz_0$. In fact, the total rate—even when nonradiative energy transfer is dominant—increases as the wavelength shrinks. From this, we conclude that for both radiative and nonradiative decay, the high confinement of 2D plasmons helps to overcome the small size of the atom.
48. C. Cohen-Tannoudji, J. Dupont-Roc, G. Grynberg, P. Thickstun, *Atom-Photon Interactions: Basic Processes and Applications* (Wiley, 1992).
49. We note that in principle the spectra increase for low frequencies because of quenching, but in the cases presented, it only provides a weak logarithmic correction to the decay rates (one to two orders of magnitude below the computed decay rates in Fig. 4F). This logarithmic correction is regulated by the energy difference between 4s and 4p levels (due to the Lamb shift). In other atoms, it is regulated by a much larger frequency (due to the lack of degeneracy between s and p orbitals), making the correction even weaker than in hydrogen (*29*).
50. A. Alù, N. Engheta, *Nat. Photonics* **2**, 307–310 (2008).


## ACKNOWLEDGMENTS


Supported by S3TEC, an Energy Frontier Research Center funded by U.S. Department of Energy grant DE-SC0001299 (B.Z. and M.S.); Marie Curie grant 328853-MC-BSiCS (I.K.); and the Army Research Office through the Institute for Soldier Nanotechnologies under contract W911NF-13-D-0001. We thank P. Rebusco for critical reading and editing of the manuscript, and H. Buljan, K. Nelson, O. Shapira, and J. Lopez for useful discussions. The authors and MIT have filed U.S. patent applications 62/342,287 and 62/266,762 that relate to the mechanisms described in this manuscript.


## SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/353/6296/263/suppl/DC1
Figs. S1 to S5
References (*51*–*55*)


6 March 2016; accepted 17 June 2016
10.1126/science.aaf6308


OZONE HOLE

# Emergence of healing in the Antarctic ozone layer


**Susan Solomon,[1]* Diane J. Ivy,[1] Doug Kinnison,[2] Michael J. Mills,[2] Ryan R. Neely III,[3,4] Anja Schmidt[3]**



**Industrial chlorofluorocarbons that cause ozone depletion have been phased out under the Montreal Protocol. A chemically driven increase in polar ozone (or "healing") is expected in response to this historic agreement. Observations and model calculations together indicate that healing of the Antarctic ozone layer has now begun to occur during the month of September. Fingerprints of September healing since 2000 include (i) increases in ozone column amounts, (ii) changes in the vertical profile of ozone concentration, and (iii) decreases in the areal extent of the ozone hole. Along with chemistry, dynamical and temperature changes have contributed to the healing but could represent feedbacks to chemistry. Volcanic eruptions have episodically interfered with healing, particularly during 2015, when a record October ozone hole occurred after the Calbuco eruption.**


Antarctic ozone depletion has been a focus of attention by scientists, policy-makers, and the public for three decades (*1*). The Antarctic ozone hole opens up in austral spring of each year and is measured both by its depth (typically a loss of about half of the total integrated column amount) and its extent (often more than 20 million km$^2$ by October). Ozone losses have also been documented in the Arctic and at mid-latitudes in both hemispheres (*2*). Concern

about ozone depletion prompted a worldwide phaseout of the production of anthropogenic halocarbons containing chlorine and bromine, which are known to be the primary sources of reactive halogens responsible for the depletion (*2*). The ozone layer is expected to recover in response, albeit very slowly, mainly because of the long atmospheric residence time of the halocarbons (*2*).

Ozone recovery involves multiple stages, starting with (i) a reduced rate of decline, followed by (ii) a leveling off of the depletion and (iii) an identifiable ozone increase that can be linked to halocarbon reductions (*2*, *3*). For simplicity, we refer to the third stage of recovery as "healing." All three stages of recovery have been documented in the upper stratosphere in mid- and low latitudes, albeit with uncertainties (*2*, *4*–*6*). Some studies provide evidence for all three recovery stages in ozone columns at mid-latitudes, despite dynamical variability (*7*). Although the first and second stages of recovery have also been well documented in the Arctic and Antarctica (*8*–*10*), a recent scientific assessment concluded that the emergence of the third stage had not been established by previous studies of the polar regions (*2*). Further, in October of 2015, the Antarctic ozone hole reached a record size (*11*), heightening questions about whether any signs of healing can be identified in either polar region.

## Controls on polar ozone

Polar ozone depletion is driven by chlorine and bromine chemistry linked to anthropogenic halocarbon emissions (*2*, *12*). But ozone is not expected to heal in a monotonic fashion as halocarbon concentrations decrease because of confounding factors (such as meteorological changes) that induce variability from one year to another and could influence trends (*2*, *13*, *14*).

The exceptionally large ozone depletion in the polar regions compared with lower latitudes is related to polar stratospheric cloud (PSC) particles that form under cold conditions. These clouds drive heterogeneous chlorine and bromine chemistry that is sensitive to small changes in temperature (and hence to meteorological variability). A second, related factor is change in the transport of ozone and other chemicals caused by circulation or mixing changes (*2*). Further, some PSCs, as well as aerosol particles capable of driving similar chemistry, are increased in abundance when volcanic eruptions increase stratospheric sulfur. Volcanically driven increases in Antarctic ozone depletion were documented in the early 1990s after the 1991 eruption of Mount Pinatubo and are well simulated by models (*15*, *16*). Since about 2005, a series of smaller-magnitude volcanic eruptions has increased stratospheric particle abundances (*17*, *18*), but the impact of these on polar ozone recovery has not previously been estimated.

[1]Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Atmospheric Chemistry Observations and Modeling Laboratory, National Center for Atmospheric Research (NCAR), Post Office Box 3000, Boulder, CO 80305, USA. [3]School of Earth and Environment, University of Leeds, Leeds, UK. [4]National Centre for Atmospheric Science, University of Leeds, Leeds, UK.
***Corresponding author. Email: solos@mit.edu**

## Observations and model test cases

We examined healing using balloon ozone data from the Syowa and South Pole stations. We also used total ozone column measurements from the South Pole station and the Solar Backscatter Ultra-Violet satellite (SBUV; we averaged SBUV data over the region from 63°S to the polar edge of coverage). The SBUV record has been carefully calibrated and compared with suborbital data (*19*). We also used the Total Ozone Mapping Spectrometer/Ozone Monitoring Instrument (TOMS/OMI) merged data set for analysis of the horizontal area of the ozone hole (*20*). Calibrated SBUV data are currently only available up to 2014, whereas the other records are available through 2015 (affecting the time intervals evaluated here). Model calculations were carried out with the Community Earth System Model 1 (CESM1) Whole Atmosphere Community Climate Model (WACCM), which is a fully coupled state-of-the-art interactive chemistry climate model (*21*). We used the specified dynamics option, SD-WACCM, in which meteorological fields including temperature and winds are derived from observations (*22*, *23*). The analysis fields allow the time-varying temperature-dependent chemistry that is important in polar ozone depletion to be simulated in detail. The model's ability to accurately represent polar ozone chemistry has recently been documented (*23*, *24*). Aerosol properties were based on the Chemistry and Climate Model Intercomparison (CCMI) recommendation (*25*) or derived inline from a version of WACCM that uses a modal aerosol submodel (*23*, *26*). The modal submodel calculates variations in stratospheric aerosols from volcanic sources (using an observational database of volcanic $SO_2$ emissions and plume altitudes; table S1) and nonvolcanic sulfur sources (particularly OCS, anthropogenic $SO_2$, and dimethyl sulfide). The injection heights and volcanic inputs are similar to those of previous studies (*18*, *27*), and the calculated aerosol distributions capture the timing of post-2005 eruptions observed in several low-, mid-, and high-latitude lidar data sets and satellite climatologies (*26*). Based on comparisons with lidar data for several eruptions and regions (*26*), our modeled post-2005 total stratospheric volcanic aerosol optical depths are estimated to be accurate to within ±40% (supplementary materials). Differences between the CCMI aerosol climatology and our calculated modal aerosol submodel results can be large, especially in the lower stratosphere (*26*), and can affect ozone abundances.

The concentrations of halogenated gases capable of depleting ozone stopped rising in the polar stratosphere around the late 1990s as a result of the Montreal Protocol and are slowly declining (*2*, *28*). We analyze what role these decreases in halogens play, along with other drivers of variability and change, in polar ozone trends since 2000. The year 2002 was anomalous in terms of meteorological behavior in the Antarctic (*29*), and it is excluded from all trend analyses throughout this paper.

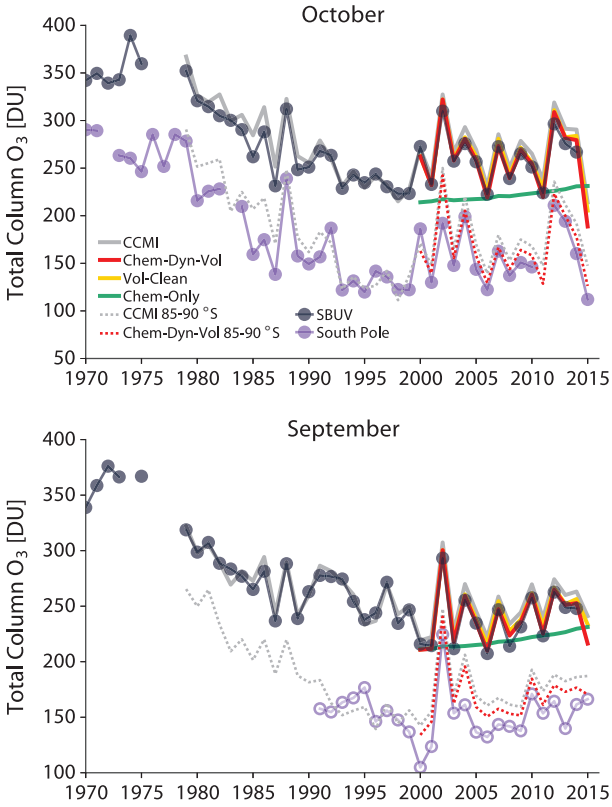



**Fig. 1. Monthly averaged Antarctic total ozone column for October and September, from SBUV and South Pole station observations and a series of model calculations.** Total ozone data measured at the geographic South Pole are from Dobson observations (filled circles) for October (**Top**) and balloon sondes (open circles) for September (**Bottom**), when there is not sufficient sunlight for the Dobson. SBUV data for each month are compared with model runs averaged over the polar cap latitude band that is accessible by the instrument; South Pole station data are compared with simulations for 85°S to 90°S.

Three different model simulations, all using full chlorine and bromine chemistry, were used to examine drivers of polar ozone changes since 2000: (i) a simulation encompassing observed time-varying changes in temperature and winds from meteorological analyses, with calculated background and volcanic stratospheric particles as well as other types of PSCs (chem-dyn-vol); (ii) a volcanically clean simulation (vol-clean), considering only background sources of stratospheric sulfur; and (iii) a chemistry-only simulation, in which annual changes in all meteorological factors (including the temperatures that drive chemistry) were suppressed by repeating conditions for 1999 throughout, and volcanically clean aerosols were imposed (chem-only). The Antarctic stratosphere in austral spring of 1999 was relatively cold and was deliberately chosen for the third simulation for its large chemical ozone losses. A longer run using full chemistry and CCMI aerosols illustrates the model's simulation of ozone loss since 1979. Further information on statistical approaches, methods, data sets, and the model are provided in the supplementary materials.

Our chem-only simulation probably represents a conservative estimate of chemical effects because it does not include radiatively driven temperature changes that are expected to occur as a result of changes in ozone (*30*) or the feedbacks of those temperature changes to chemical processes. Temperature and ozone are coupled because absorption of sunlight by ozone heats the stratosphere. If ozone increases as a result of reductions in halogens, then temperatures will increase, which feeds back to the chemistry (for example, by reducing the rate of temperature-dependent heterogeneous reactions that deplete ozone), further increasing ozone. Such effects have not been separated here from other changes in temperature or winds (due to dynamical variability or forcings such as greenhouse gases).

## Antarctic ozone trends, variability, and fingerprints of healing

Most analyses of Antarctic ozone recovery to date consider October or September-to-November averages (*7*, *9*, *10*). The historic discovery of the Antarctic ozone hole was based on observations taken in October (*1*), and healing cannot be considered complete until the ozone hole ceases to occur in that month, which is expected around mid-century (*2*, *28*). However, October need not be the month when the onset of the healing process occurs. A first step in understanding whether a signal of the onset of healing can be identified is examining trends and their statistical significance relative to the noise of interannual variability.

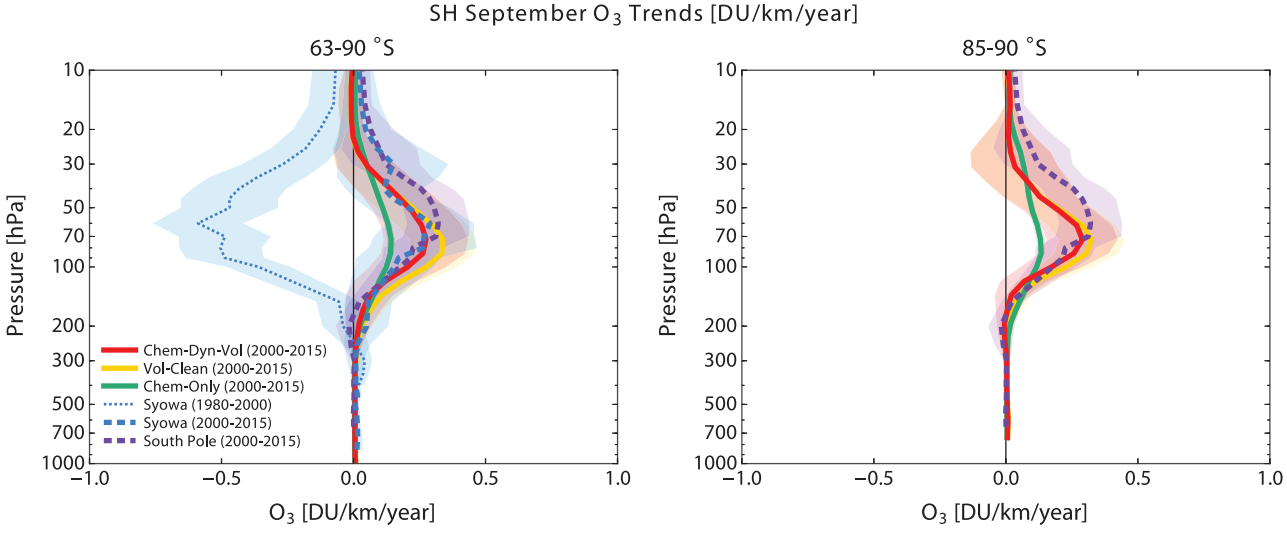



**Fig. 2. Trends in Southern Hemisphere (SH) polar cap ozone profiles in September.** Ozone data from balloons at the Syowa (69°S, 39.58°E) (**Left**) and South Pole (**Right**) stations, along with model simulations averaged over the polar cap and over 85°S to 90°S, respectively, are shown versus pressure. The shading represents the uncertainties on the trends at the 90% statistical confidence interval.

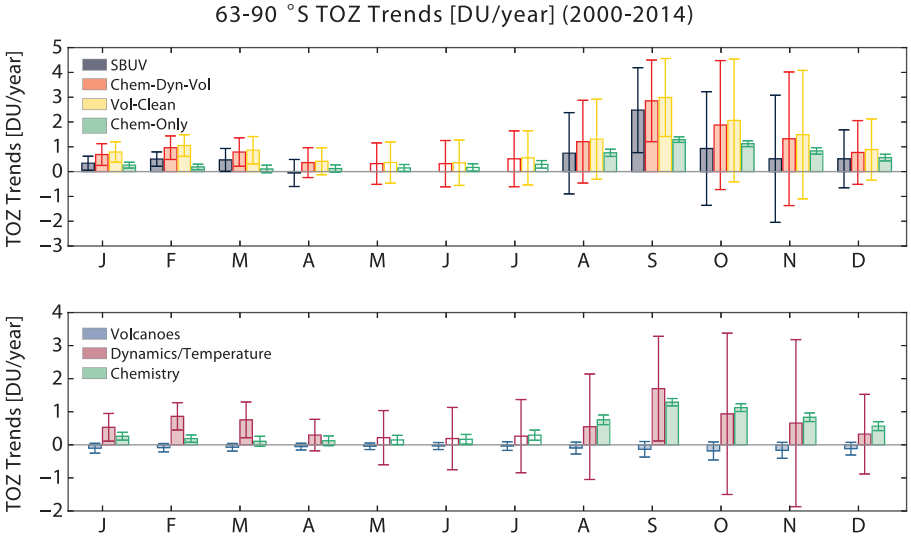



**Fig. 3. Trends in total ozone abundance (TOZ) by month from 2000 to 2014.** (**Top**) Monthly and polar cap–averaged SBUV satellite observations, together with model simulations masked to the satellite coverage. Error bars denote 90% statistical confidence intervals. (**Bottom**) Contributions to the simulated monthly trends in total ozone abundance driven by dynamics and temperature (vol-clean minus chem-only), chemistry only (chem-only), and volcanoes (chem-dyn-vol minus vol-clean). In austral winter, SBUV measurements do not extend to 63°S; therefore, the model averages for those months cover 63°S to 90°S (open bars).

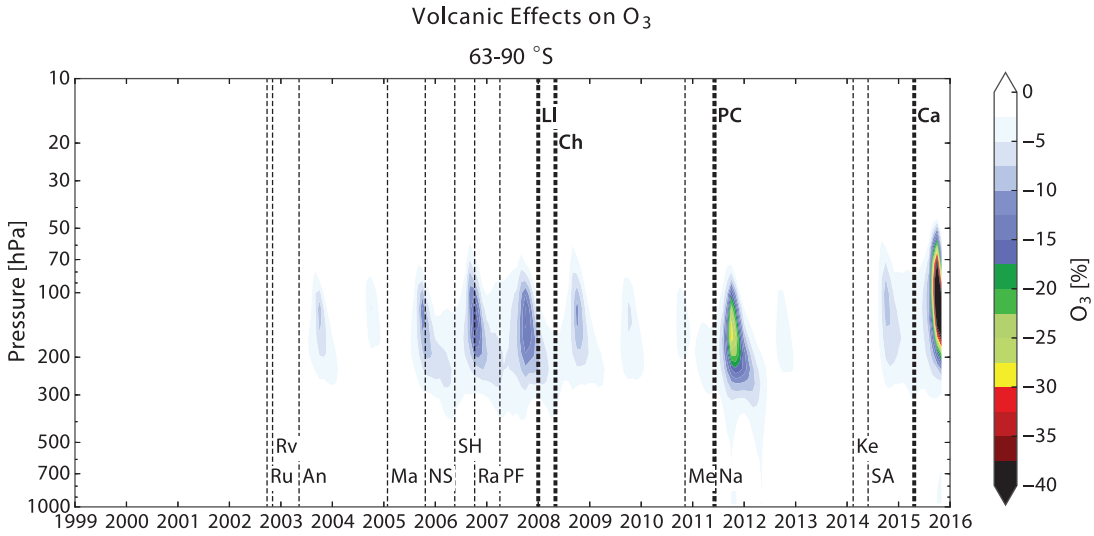



**Fig. 4. Model-calculated percentage changes in local concentrations of ozone due to a series of moderate volcanic eruptions.** The results (chem-dyn-vol minus vol-clean simulations) are averaged over the Antarctic polar cap as a function of pressure and month. Volcanic eruptions that have dominated the changes are indicated, with tropical eruptions at the bottom and higher-latitude eruptions at the top. An, Anatahan; Ca, Calbuco; Ch, Chaiten; Ke, Kelut; Ll, Llaima; Ma, Manam; Me, Merapi; Na, Nabro; NS, Negra Sierra; PC, Puyehue-Cordón Caulle; PF, Piton de la Fournaise; Ra, Rabaul (also referred to as Tavurvur); Ru, Ruang; Rv, Reventador; SA, Sangeang Api; SH, Soufriere Hills.

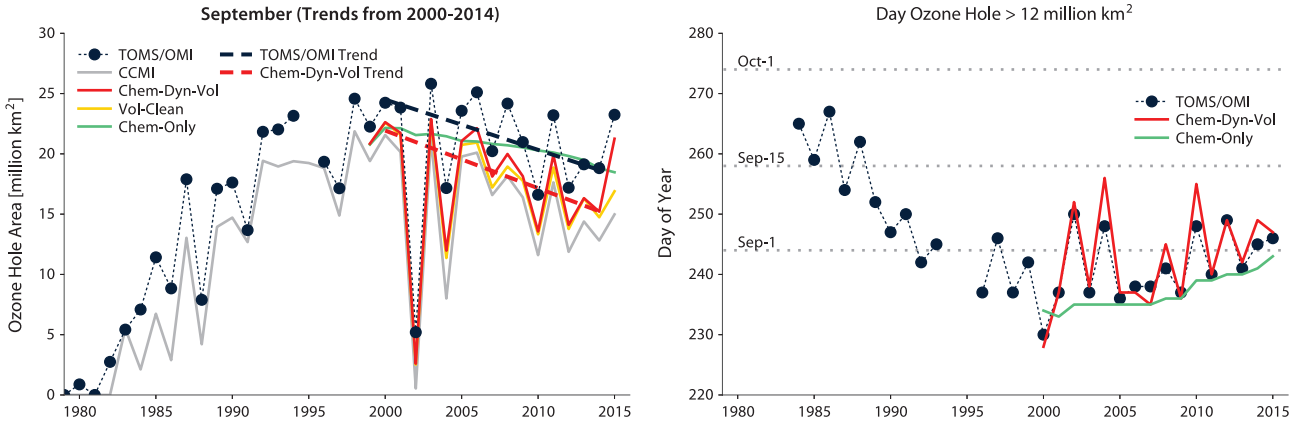



**Fig. 5. September size of the ozone hole and day of year when the hole exceeds 12 million km$^2$** (**Left**) Annual size of the September monthly averaged ozone hole, defined as the region where the total ozone amount is less than 220 DU, from TOMS/OMI satellite observations and model simulations. Trends in the TOMS/OMI observations (heavy dashed black line) and the chem-dyn-vol model calculations (heavy dashed red line) from 2000 to 2015 are also indicated. (**Right**) Annual day of the year when the size of the ozone hole exceeds 12 million km$^2$ (and remains above that value for at least 3 days) in the TOMS/OMI observations and model simulations.

October has the deepest ozone depletion of any month in the Antarctic. However, it is subject to large variability due to seasonal fluctuations in temperature and transport, as well as to volcanic aerosol chemistry. Figure 1 shows the time series of measured Antarctic total ozone in October obtained from SBUV and South Pole station data, along with the model calculations; tables S2 and S3 provide the associated post-2000 trends and 90% confidence intervals. SD-WACCM reproduces the observed October variability from year to year when all factors are considered (chem-dyn-vol) (Fig. 1). However, the October total ozone trends are not yet positive with 90% certainty in the observational data or in the model. In contrast, other months with lower depletion and variability (particularly September; Fig. 1, fig. S1, and tables S2 and S3) reveal positive ozone trends from 2000 to 2014 that are statistically significant at 90% confidence in SBUV and South Pole station measurements. Arctic ozone has long been known to be more variable than Antarctic ozone (*2*), and no Arctic month yet reveals a significant positive trend in either the chem-dyn-vol simulation or the SBUV observations when examined in the same manner (table S2).

The September profile of balloon ozone trends is a key test of process understanding. Figure 2 shows measured balloon profile trends for the South Pole and Syowa stations for 2000–2015, together with SD-WACCM model simulations. The large ozone losses measured at Syowa as the ozone hole developed from 1980 to 2000 are also shown for comparison. Antarctic station data need to be interpreted with caution because of an observed long-term shift in the position of the Antarctic vortex that affects Syowa in particular in October; the South Pole station is, however, less influenced by this effect (*31*). The ozonesonde data sets suggest clear increases since 2000 between about 100 and 50 hPa (*10*). The chem-only simulation reproduces about half of the observed healing, with the remainder in this month being accounted for by dynamics and temperature. The simulations also suggest a negative contribution (offset to healing) due to volcanic enhancements of the ozone depletion chemistry between about 70 and 200 hPa (fig. S2 shows similar effects in other months in this sensitive height range). The comparisons of the modeled trend profiles in Fig. 2 provide an important fingerprint of the onset of healing of

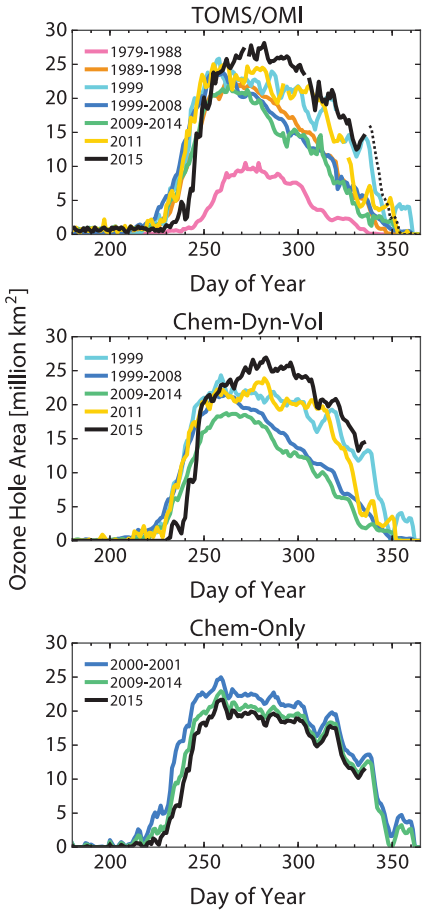



**Fig. 6. Daily size of the ozone hole.** Daily measurements (**Top**) and model calculations (**Middle** and **Bottom**) of the size of the Antarctic ozone hole versus day of year for different time intervals, with 2015 shown in black. The dashed black line in the top panel denotes the 2015 TOMS/OMI data after the period covered by the model runs.

the Antarctic ozone layer in September. This is consistent with the basic understanding that reductions in ozone-depleting substances in the troposphere will lead to a healing of polar ozone that emerges over time, with lags due to the transport time from the troposphere to the stratosphere, along with the time required for chemically driven trends to become significant relative to dynamical and volcanic variability.

The seasonal cycle of monthly total ozone trends from the SBUV satellite observations is shown in Fig. 3, along with model calculations for various cases. The contributions to the modeled trends that are due to volcanic inputs (difference between chem-dyn-vol and vol-clean simulations), chemistry alone (chem-only simulations), and dynamics and temperature (difference between vol-clean and chem-only simulations) are shown in the lower panel. Although it is not possible to be certain that the SBUV observations vary for the same reasons that the model calculations do, the broad agreement between them on the seasonal cycle of total trends supports this interpretation. Less dynamical variability in September compared with October (as shown by smaller error bars on the "dynamics/temperature" term in Fig. 3, bottom panel), along with strong chemical recovery, make September the month when the Antarctic ozone layer has undergone the largest amount of healing since 2000. The data suggest September increases (at 90% confidence) of 2.5 ± 1.6 DU per year over the latitudes sampled by SBUV and 2.5 ± 1.5 DU per year from the South Pole sondes. These values are consistent with the chem-dyn-vol model values of 2.8 ± 1.6 and 1.9 ± 1.5 DU per year, respectively. Because the model reproduces much of the observed year-to-year variability in September total ozone from both the South Pole station and SBUV observations, confidence is increased that there is a significant chemical contribution to the trends (Fig. 1). As a best estimate, the model results suggest that roughly half of the September column healing is chemical, and half is due to dynamics and temperature, though the dynamic and temperature effects are highly variable. The modeled total September healing trend has been reduced by about 10% as a result of the chemical effects of increased volcanic activity in the latter part of 2000–2014.

Volcanic eruptions affect polar ozone depletion because injections of sulfur increase the surface areas of liquid PSCs and aerosol particles (*32*). Higher-latitude eruptions directly influence the polar stratosphere, but tropical eruptions also can increase polar aerosols via transport. The model indicates that numerous moderate eruptions since about 2005 have affected polar ozone in both hemispheres (table S1 gives eruptions, dates, and latitudes), particularly at pressures from about 70 to 300 hPa (Fig. 4). At pressures above about 100 hPa, temperatures are generally too warm for many PSCs to form, but there is sufficient water that effective heterogeneous chemistry can take place under cold polar conditions (*12*). Our simulations indicate peak volcanic losses locally as large as 30 and 55% for the Antarctic in 2011 and 2015, mainly due to the Chilean eruptions of Puyehue-Cordón Caulle and Calbuco, respectively; contributions to depletions that are attributable to tropical eruptions are also evident in several earlier years. At these pressures, contributions to the total column are small but significant: The integrated additional Antarctic ozone column losses, averaged over the polar cap, are between 5 and 13 DU after the eruptions shown in Fig. 4.

The ozone hole typically begins to open in August of each year and reaches its maximum areal extent in October. Decreases in the areal extent of the October hole are expected to occur in the 21st century as chemical destruction slows, but they cannot yet be observed against the backdrop of interannual variability, in part because of the extremely large hole in 2015 (fig. S3). However, monthly averaged observations for September show a shrinkage of 4.5 ± 4.1 million km$^2$ between 2000 and 2015 (Fig. 5, left panel). The model underestimates the observed September hole size by about 15% on average, but it yields variability (Fig. 5) and trends (4.9 ± 4.7 million km$^2$) that are similar to the observations. The right panel of Fig. 5 shows that the observed and modeled day of the year when the ozone hole exceeds a threshold value of 12 million km$^2$ has been occurring later in recent years, indicating that early September holes are becoming smaller (Fig. 6). This result is robust to the specific choice of threshold value and implies that the hole is opening more slowly as the ozone layer heals. The chem-only simulation results in Fig. 5 show that if temperatures, dynamical conditions, and volcanic inputs had remained the same as they were in 1999 until now, the September ozone hole would have shrunk by about 3.5 ± 0.3 million km$^2$ as a result of reduced chlorine and bromine; thus, the reduction of these chemicals dominates the total shrinkage over this period.

Volcanic eruptions caused the modeled areas of the September-averaged ozone holes to expand substantially in several recent years. Our results, as shown in Fig. 5 (left panel), indicate that much of the statistical uncertainty in the observed September trend is not random, but rather is due to the expected chemical impacts of these geophysical events. In 2006, 2007, and 2008, model calculations suggest that the September ozone holes were volcanically enlarged by about 1 million km$^2$. The sizes of the September ozone holes in 2011 and 2015 are estimated to have been, respectively, about 1.0 million and 4.4 million km$^2$ larger because of volcanic eruptions (especially Puyehue-Cordón Caulle in 2011 and Calbuco in 2015) than they would otherwise have been, substantially offsetting the chemical healing in those years.

Figure 6 shows that the bulk of the seasonal growth of the ozone hole typically occurs between about days 230 and 250 (late August to early September). As the ozone layer heals, the growth of the hole is expected to occur later in the year (middle and bottom panels), in agreement with observations (top). The slower rates of early-season growth are key to the trend of shrinkage in the September-averaged ozone hole. For example, the rate of ozone loss depends strongly upon the ClO concentration, so that reduced chlorine concentrations lead to slower rates of ozone loss after polar sunrise. The ozone hole of 2015 was considerably larger over several weeks in October (but notably, not in September) than ever previously observed, and this behavior is well reproduced in our model only when the eruption of Calbuco is considered (figs. S3 and S4). The record-large monthly averaged ozone hole in October 2015 measured 25.3 million km$^2$, which is 4.8 million km$^2$ larger than the previous record year (20.6 million km$^2$ in 2011). When volcanic aerosols are included (chem-dyn-vol simulation), our calculated monthly averaged October 2015 ozone hole is 24.6 million km$^2$, whereas the corresponding value excluding volcanic aerosols (vol-clean simulation) is much smaller, 21.1 million km$^2$ (fig. S3). Therefore, our calculations indicate that cold temperatures and dynamics alone made a much smaller contribution to establishing the October 2015 record than volcanic aerosols did (figs. S3 and S4), and the cold temperatures are expected to be at least partly a feedback to the volcanically increased large ozone losses. Further, the conclusion that the volcanic aerosols were the dominant cause of the

record size of the October 2015 ozone hole would hold based on our calculations even if the volcanic aerosol amounts were overestimated by a factor of several (a much larger error than indicated by our comparison of our model calculations with lidar data for multiple eruptions in (*26*); supplementary materials).

The reasons for the contributions of dynamics and temperature to the healing of the Antarctic ozone layer are not clear. The dynamical and temperature contributions to healing estimated in Fig. 3 vary by month in a manner that mirrors the ozone depletion in spring, suggesting linkages to the seasonality of the depletion itself and hence possible dynamical feedbacks. Some models (*33*–*35*) suggest that a reduction in transport of ozone to the Antarctic occurred as depletion developed in the 1980s and 1990s, which would imply a reversal of this process and hence increased healing as ozone rebounds. But others indicate that ozone depletion increased the strength of the stratospheric overturning circulation (*36*), and a reversal of this factor during recovery would impede healing. Although there is robust agreement across models that climate change linked to increasing greenhouse gases should act to increase the strength of the stratospheric overturning circulation, observations show mixed results (*37*); further, the seasonality of this effect has not been established, and the magnitude in the Antarctic is uncertain. Internal variability of the climate system linked, for example, to variations in El Niño could also affect the trends.

## Conclusion

After accounting for dynamics, temperature, and volcanic factors, the results presented here indicate that healing of the Antarctic ozone hole is emerging. Our results underscore the combined value of balloon and satellite ozone data, volcanic aerosol measurements, and chemistry-climate models in documenting progress in the recovery of the ozone layer since the Montreal Protocol.

### REFERENCES AND NOTES


1. J. C. Farman, B. G. Gardiner, J. D. Shanklin, *Nature* **315**, 207–210 (1985).
2. World Meteorological Organization/United Nations Environment Programme (WMO/UNEP), *Scientific Assessment of Ozone Depletion: 2014* (Global Ozone Research and Monitoring Project Report No. 55, WMO, 2014).
3. D. J. Hofmann, S. J. Oltmans, J. M. Harris, B. J. Johnson, J. A. Lathrop, *J. Geophys. Res.* **102**, 8931–8943 (1997).
4. M. J. Newchurch *et al.*, *J. Geophys. Res.* **108**, 4507 (2003).
5. N. R. P. Harris *et al.*, *Atmos. Chem. Phys.* **15**, 9965–9982 (2015).
6. F. Tummon *et al.*, *Atmos. Chem. Phys.* **15**, 3021–3043 (2015).
7. T. G. Shepherd *et al.*, *Nat. Geosci.* **7**, 443–449 (2014).
8. E.-S. Yang *et al.*, *J. Geophys. Res.* **113**, D20308 (2008).
9. M. L. Salby, E. Titova, L. Deschamps, *Geophys. Res. Lett.* **38**, L09702 (2011).
10. J. Kuttippurath *et al.*, *Atmos. Chem. Phys.* **13**, 1625–1635 (2013).
11. WMO, "WMO Antarctic Ozone Bulletins: 2015," 2015; www.wmo.int/pages/prog/arep/WMOAntarcticOzoneBulletins2015.html.
12. S. Solomon, *Rev. Geophys.* **37**, 275–316 (1999).
13. J. Kuttippurath *et al.*, *Atmos. Chem. Phys.* **15**, 10385–10397 (2015).
14. N. J. Livesey, M. L. Santee, G. L. Manney, *Atmos. Chem. Phys.* **15**, 9945–9963 (2015).
15. D. J. Hofmann, S. J. Oltmans, *J. Geophys. Res.* **98**, 18555–18561 (1993).
16. R. W. Portmann *et al.*, *J. Geophys. Res.* **101**, 22991–23006 (1996).
17. J.-P. Vernier *et al.*, *Geophys. Res. Lett.* **38**, L12807 (2011).
18. C. Brühl, J. Lelieveld, H. Tost, M. Höpfner, N. Glatthor, *J. Geophys. Res. Atmos.* **120**, 2103–2118 (2015).
19. R. D. McPeters, P. K. Bhartia, D. Haffner, G. L. Labow, L. Flynn, *J. Geophys. Res.* **118**, 8032–8039 (2013).
20. W. Chehade, M. Weber, J. P. Burrows, *Atmos. Chem. Phys.* **14**, 7059–7074 (2014).
21. D. R. Marsh *et al.*, *J. Clim.* **26**, 7372–7391 (2013).
22. A. Kunz, L. L. Pan, P. Konopka, D. E. Kinnison, S. Tilmes, *J. Geophys. Res.* **116**, D24302 (2011).
23. Materials and methods are available as supplementary materials on *Science* Online.
24. S. Solomon, D. Kinnison, J. Bandoro, R. R. Garcia, *J. Geophys. Res.* **120**, 7958–7974 (2015).
25. F. Arfeuille *et al.*, *Atmos. Chem. Phys.* **13**, 11221–11234 (2013).
26. M. J. Mills *et al.*, *J. Geophys. Res.* **121**, 2332–2348 (2015).
27. M. Höpfner *et al.*, *Atmos. Chem. Phys.* **15**, 7017–7037 (2015).
28. P. A. Newman, E. R. Nash, S. R. Kawa, S. A. Montzka, S. M. Schauffler, *Geophys. Res. Lett.* **33**, L12814 (2006).
29. A. A. Scaife *et al.*, *J. Atmos. Sci.* **62**, 629–639 (2005).
30. P. M. Forster, R. S. Freckleton, K. P. Shine, *Clim. Dyn.* **13**, 547–560 (1997).
31. B. Hassler, G. E. Bodeker, S. Solomon, P. J. Young, *Geophys. Res. Lett.* **38**, L01805 (2011).
32. A. Tabazadeh, K. Drdla, M. R. Schoeberl, P. Hamill, O. B. Toon, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 2609–2612 (2002).
33. S. Meul, S. Oberländer-Hayn, J. Abalichin, U. Langematz, *Atmos. Chem. Phys.* **15**, 6897–6911 (2015).
34. P. Braesicke *et al.*, *Atmos. Chem. Phys.* **13**, 10677–10688 (2013).
35. F. Li, J. Austin, J. Wilson, *J. Clim.* **21**, 40–57 (2008).
36. C. McLandress, T. G. Shepherd, *J. Clim.* **22**, 1516–1540 (2009).
37. N. Butchart, *Rev. Geophys.* **52**, 157–184 (2014).


### ACKNOWLEDGMENTS

We thank S. Tilmes (NCAR) for help with the MERRA (Modern-Era Retrospective Analysis for Research and Applications) data. D.K. and S.S. were partially supported by NSF Frontiers in Earth System Dynamics grant OCE-1338814, and D.J.I. was supported by NSF Atmospheric Chemistry grant AGS-1539972. A.S. was supported by an Academic Research Fellowship from the University of Leeds, an NCAR visiting scientist grant, and Natural Environment Research Council grant NE/N006038/1. NCAR is sponsored by NSF. WACCM is a component of CESM, which is supported by NSF and the Office of Science of the U.S. Department of Energy. We are grateful to D. Fahey, B. Hassler, W. D. McKenna, and the anonymous reviewers for helpful comments. Instructions for access to data reported in this paper are given in the supplementary materials.

### SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/353/6296/269/suppl/DC1
Materials and Methods
Figs. S1 to S4
Tables S1 to S3
References (*38*–*44*)

6 December 2015; accepted 20 June 2016
Published online 30 June 2016
10.1126/science.aae0061



REPORTS

FERROELECTRICITY

# Discovery of robust in-plane ferroelectricity in atomic-thick SnTe

Kai Chang,[1,2]* Junwei Liu,[3,1,2]* Haicheng Lin,[1,2] Na Wang,[1,2] Kun Zhao,[1,2] Anmin Zhang,[4] Feng Jin,[4] Yong Zhong,[1,2] Xiaopeng Hu,[1,2] Wenhui Duan,[1,2] Qingming Zhang,[4,5] Liang Fu,[3] Qi-Kun Xue,[1,2] Xi Chen,[1,2]† Shuai-Hua Ji[1,2,6]†

**Stable ferroelectricity with high transition temperature in nanostructures is needed for miniaturizing ferroelectric devices. Here, we report the discovery of the stable in-plane spontaneous polarization in atomic-thick tin telluride (SnTe), down to a 1–unit cell (UC) limit. The ferroelectric transition temperature $T_c$ of 1-UC SnTe film is greatly enhanced from the bulk value of 98 kelvin and reaches as high as 270 kelvin. Moreover, 2- to 4-UC SnTe films show robust ferroelectricity at room temperature. The interplay between semiconducting properties and ferroelectricity in this two-dimensional material may enable a wide range of applications in nonvolatile high-density memories, nanosensors, and electronics.**

Two-dimensional (2D) materials exhibit a wide range of symmetry-breaking quantum phenomena such as crystalline order (*1*, *2*), superconductivity (*3*, *4*), magnetism (*5*, *6*), and charge-density wave (*7*, *8*), which persist in the limit of a single-unit-cell thickness. It has been comparatively more difficult to explore ferroelectricity in ultrathin films. For the perovskite ferroelectric materials, stable out-plane spontaneous polarization has been discovered in the thin films of few unit cells (UCs) thickness (*9*–*12*). Theoretical studies point out that the charge screening, chemical bonding, and misfit strain at the interface may play a role in stabilizing ferroelectric states (*13*–*18*). The transition temperature in those ultrathin films usually decreases as the thickness is reduced (*10*–*12*, *19*), which could be understood by the destabilization of ferroelectric states through depolarization fields. In contrast, the transition temperature of ferroelectric polymer film is nearly independent of thickness, indicating 2D behavior (*20*).

Here, we use the molecular beam epitaxial technique to prepare atomic-thick ferroelectric

SnTe films and discover stable in-plane spontaneous polarization in SnTe film of only one UC thickness, 0.63 nm. To avoid external strain effect, we employ the graphitized 6H-SiC(0001) substrate in epitaxial growth to obtain stress-free thin films. The substrate surface is mainly covered by the monolayer/bilayer graphene. The weak van der Waals bonding between substrate and SnTe film largely reduces the strain effect and helps to preserve the in-plane component of the polarization.

[1]State Key Laboratory of Low-Dimensional Quantum Physics, Department of Physics, Tsinghua University, Beijing 100084, China. [2]Collaborative Innovation Center of Quantum Matter, Beijing 100084, China. [3]Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [4]Department of Physics, Beijing Key Laboratory of Optoelectronic Functional Materials and Micro-Nano Devices, Renmin University of China, Beijing 100872, China. [5]Collaborative Innovation Center of Advanced Microstructures, Nanjing 210093, China. [6]RIKEN Center for Emergent Matter Science (CEMS), Wako, Saitama 351-0198, Japan.
*These authors contributed equally to this work. **†Corresponding author. Email: xc@mail.tsinghua.edu.cn (X.C.); shji@mail.tsinghua.edu.cn (S.-H.J.)**

In the bulk form, SnTe is a narrow-gap (~0.2 eV) semiconductor (*21*) and possesses the rock-salt structure with a lattice constant of 6.32 Å (Fig. 1A) at room temperature. SnTe is always heavily p-type doped because of the negative formation energy of Sn vacancy (*22*). At the ferroelectric transition temperature $T_c$, the crystal goes through a cubic-to-rhombohedral structural phase transition and the two sublattices of Sn and Te atoms are displaced from each other along the [111] direction, forming the ferroelectric state (*23*, *24*). Owing to the screening effect of charge carriers, $T_c$ drops rapidly as the concentration of Sn vacancy increases (*25*), and the highest transition temperature of bulk SnTe is only 98 K (*26*). Recently, interest in the ferroelectric structural distortion in SnTe was rekindled by its effect on electronic properties of the topological crystalline insulator phase newly discovered in group IV-VI semiconductors (*27*).

The SnTe films grown on SiC are characterized by the in situ reflection high-energy electron diffraction (RHEED) [(*28*), fig. S1] and scanning tunneling microscopy (STM) topography images (Fig. 1B and fig. S1). At low coverage, (001)-oriented islands are formed on the substrate. In contrast to the case involving strong directional covalent bonds, the lattice-matching conditions are largely relaxed in the van der Waals epitaxy. As a consequence, the in-plane orientation of the SnTe islands is randomly distributed. The weak bonding is also demonstrated by the fact that an entire island at the size of 100 nm can be displaced by the STM tip (fig. S2). By carefully controlling the substrate temperature and SnTe flux, the size of an island can be as large as ~1 μm (fig. S1). The facet edges along the [100] and [010] directions are clearly seen in the STM images. The island thickness corresponds to an integer multiple of the SnTe unit cell (no half layer), which is consistent with the formation energy calculation (fig. S3). The topography images (fig. S4) also indicate that the formation of Sn vacancies is greatly suppressed in the atomic-thin SnTe films. The surface defect density is $10^{10}$ to $10^{11}$ $cm^{-2}$ for the 2-UC film and even lower for

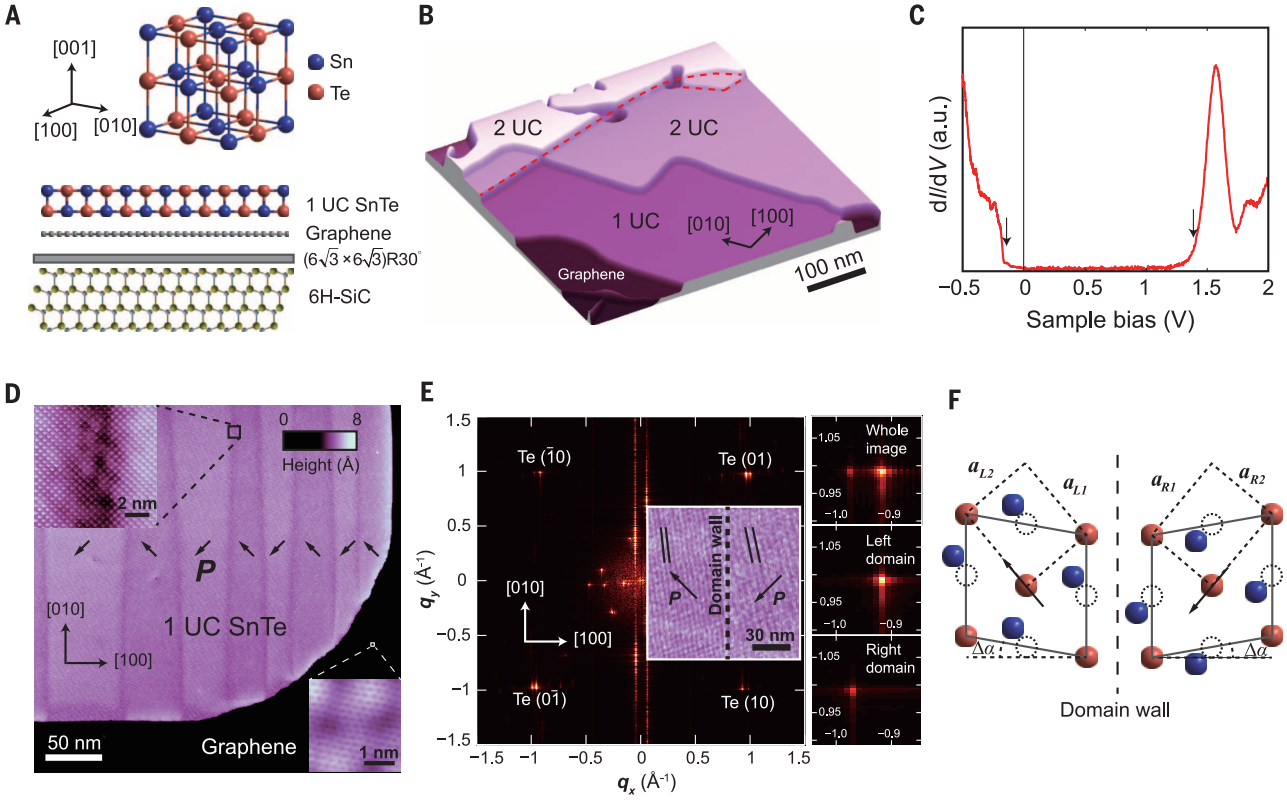



**Fig. 1. Structure and lattice distortion of SnTe film.** (**A**) Schematics of the SnTe crystal structure (upper) and the SnTe film (lower). (**B**) Typical STM topography image of SnTe film (sample bias 3.0 V, tunneling current 30 pA, temperature 4.7 K). The red dotted line indicates the steps of substrate. (**C**) The d*I*/d*V* spectra acquired on the surface of a 1-UC film at 4.7 K. The d*I*/d*V* of conduction and valence bands has large difference in intensity. For clarity, the spectra above and below the Fermi level are measured under different tunneling conditions: 3.0 V, 100 pA above and −0.5 V, 100 pA below. The arrows indicate the edges of the valence and conduction bands. The peak at 1.5 V corresponds to a van Hove singularity in the conduction band. (**D**) The stripe domain of a 1-UC SnTe film (imaging conditions: −0.2 V, 30 pA, 4.7 K). The arrows in each domain indicate the direction of lattice distortion. (Upper inset) Topography image across a domain boundary (−0.2 V, 100 pA). (Lower inset) The graphene substrate (−0.1 V, 200 pA). (**E**) Fourier transform (left) of an area (inset, 4.7 K) crossing a domain boundary. The Bragg peaks are associated with the Te sublattice. The parallel lines in the inset indicate the moiré stripes in each domain. The atomically resolved image right on the domain boundary is shown in fig. S5. The Te($\bar{1}$0) peaks for the whole image, the left domain, and the right domain are enlarged in the right panels. (**F**) Schematic of the lattice distortion and atom displacement in the ferroelectric phase. The solid lines indicate the rock-salt unit cell, and the dashed lines indicate the primitive cell of the Te sublattice. The arrows point to the directions of distortion. $a_{L1}$, $a_{L2}$, $a_{R1}$, and $a_{R2}$ are the lattice constants of the Te sublattice, and $\Delta\alpha$ is the distortion angle of the rock-salt unit cell. For more details, see figs. S6 to S9.

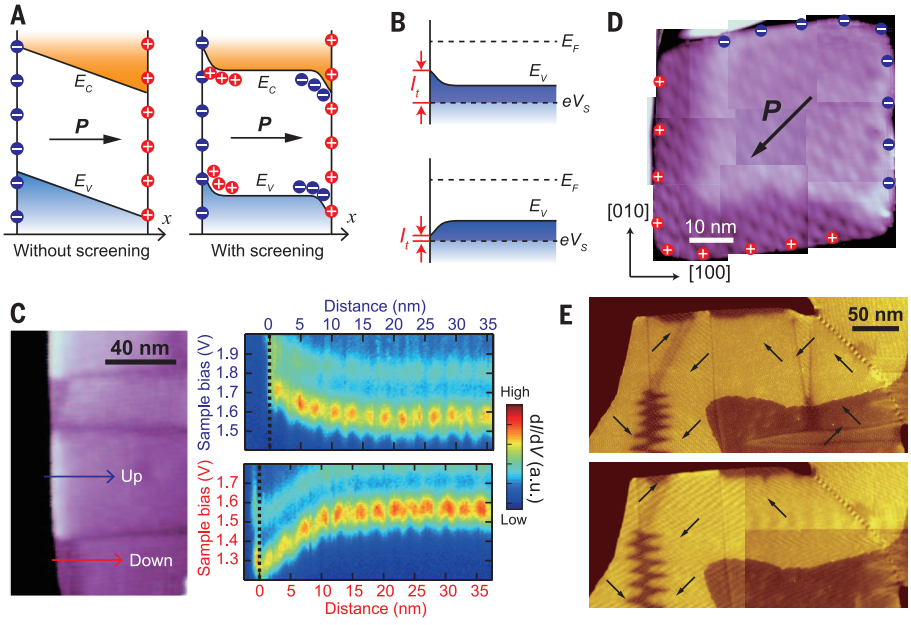



**Fig. 2. Band-bending induced by spontaneous polarization and manipulation of the domain walls.** (**A**) Schematics showing the band shift of a bulk ferroelectric with and without internal screening charge. (**B**) The apparent height of an edge. If an edge is negatively charged (top panel), more states between Fermi level ($E_F$) and the bias ($eV_s$) are available for tunneling. Therefore, the STM tip has to be lifted up to keep a constant current during scanning, giving rise to a higher apparent height. A positively charged edge has the opposite trend. (**C**) Spatially resolved d*I*/d*V* spectra (right panels) obtained along the two arrows in the image on the left (1 UC, 3.0 V, 100 pA, 4.7 K). A few typical d*I*/d*V* curves (near both conduction and valence band edges) and the fitting of the bending profiles are shown in figs. S11 and S12. (**D**) STM image of a 1-UC single-domain island (−0.2 V, 100 pA, 4.7 K). The polarization is unambiguously determined to be along the [$\bar{1}\bar{1}$0] direction. The image is a combination of nine smaller scans. The "+" and "−" signs on the edges indicate the positive and negative polarization charges. (**E**) Topography images (−0.2 V, 50 pA, 4.7 K) of the same area before (upper) and after (lower) a 5-V voltage pulse is applied for 50 ms on the 1-UC film. The location where the pulse is applied is shown in fig. S16. The arrows indicate the direction of polarization.

the 1-UC one. Both are much lower than that of the bulk material (~$10^{13}$ $cm^{-2}$). The electronic density of states of 1-UC SnTe film is measured by scanning tunneling spectroscopy (STS). The arrows in the differential conductance (d*I*/d*V*) curve (Fig. 1C) indicate the band edges. The energy gap increases from its bulk value of 0.2 to 1.6 eV in a 1-UC film, mainly originating from the quantum confinement (figs. S23 and S24).

It is challenging to probe the ferroelectric properties of an ultrathin film because of the much-reduced signal compared with the bulk. Usually, more sensitive probes, such as the synchrotron x-ray scattering (*10*), ultraviolet Raman spectroscopy (*12*), polarized second-harmonic generation measurement (*29*), and piezoresponse force microscope (*30*), are employed to detect such weak signals. In our experiment, the small island size requires even more sensitive measurement with high spatial resolution. Here, we use STM and STS to probe the ferroelectricity of SnTe in the 2D limit. The observed evidence for ferroelectricity includes the formation of domain structure (Fig. 1D), lattice distortion (Fig. 1, E and F), band-bending (Fig. 2, A to D), and polarization manipulation by electric field (Fig. 2E).

The domain structure in 1-UC SnTe film is resolved in the STM image (Fig. 1D) at certain sample bias voltage (for example, −0.2 V). The parallel stripes are along the [010] direction. The atomically resolved image (Fig. 1D, left inset) shows a quasi-square lattice of Te atoms (the negative bias corresponds to the filled states mainly contributed by Te). The lattice constant ~ 4.5 Å is in good agreement with the Te-Te distance in the (001) plane of bulk SnTe. The lattice is continuous across the domain boundary (Fig. 1D, left inset). Nevertheless, the Fourier transform of an area containing two domains (STM image in the inset of Fig. 1E) clearly exhibits two sets of Bragg peaks split along the [100] direction (Fig. 1E). Each set of the Bragg peaks is contributed by one domain (see the right panels of Fig. 1E). The lattice is slightly distorted from a perfect square to a parallelogram (Fig. 1F). From the Fourier transform, the two lattice constants of the Te sublattice at liquid helium temperature are found to be 4.58 ± 0.05 and 4.44 ± 0.05 Å, respectively. The parallelogram is elongated along the [110] and its equivalent orientations (indicated by arrows in Fig. 1, D to F). The elongated diagonals for two adjacent domains are perpendicular to each other. As shown later, the in-plane polarization is along those diagonals.

The domain formation and lattice distortion are still not adequate to serve as the conclusive evidence for ferroelectricity. A more direct manifestation comes from the band-bending at the edge of an island. In general, the discontinuity of polarization on the border of a dielectric induces a polarization charge, the density of which is given by $\vec{P} \cdot \vec{n}$. Here, $\vec{P}$ and $\vec{n}$ are the polarization and the normal direction, respectively. If one surface is positively charged, the opposite surface must be negatively charged. The electric field generated by the polarization charge shifts the bulk electronic bands (Fig. 2A). If there are free carriers owing to doping, the screening effect confines the band-bending within the vicinity of the borders (Fig. 2A). Band-bending has various origins and is commonly observed on surfaces. However, the unique feature on the surfaces of ferroelectrics is that the bending directions are opposite to each other on the opposite surfaces of a domain; one side is upward, and the other side is downward (*31*).

The signatures of band-bending can be seen in Fig. 1D. In the STM image (constant current mode at −0.2 V), the height of the island edge is different from that of the bulk, and the edges for two adjacent domains show opposite variation in height. The observed pattern can be easily explained by the band-bending effect if the polarization direction is indeed given by the arrows in Fig. 1D (see also fig. S10). The band gap at an edge is shifted by the polarization charge (Fig. 2B). The positive charge moves the band downward, and there is less density of states between the bias voltage $V_s$ and the Fermi level. Therefore, the apparent height of a positively charged edge is lower than that of the bulk. Similarly, a negatively charged edge appears higher.

The band-bending is even more clearly observed by following the peak at 1.5 V in the d*I*/d*V* curves as a function of the distance to an edge (Fig. 2C). Spatially resolved d*I*/d*V* spectra (right panels) are taken along the lines perpendicular to the edges of two adjacent domains (arrows in the left panel). The peaks shift to opposite directions up to 0.2 eV with a screening length of about 10 nm. For comparison, the d*I*/d*V* mapping is also performed on a 1-UC PbTe island and shows no band-bending effect (fig. S13). PbTe is paraelectric but otherwise is very similar to SnTe.

The above observations do not uniquely determine the orientation of polarization. Any configuration of polarization gives rise to a similar band-bending pattern if the projections of polarization on the normal direction of edge are opposite to each other between two adjacent domains in Fig. 1D. The polarization orientation is unequivocally determined by the band-bending pattern on a single domain island (Fig. 2D). The reversed bending on the opposite edges of the square island clearly demonstrates that the polarization of 1-UC SnTe film has in-plane component along the [110] diagonal.

With the in-plane polarization determined by lattice distortion together with the sign change of polarization charge on edges, we are able to classify the different types of domain walls. The straight domain walls in Fig. 1D belong to the 90° "head-to-tail" type. The "head-to-head" domain wall in Fig. 2E shows the zigzag pattern to

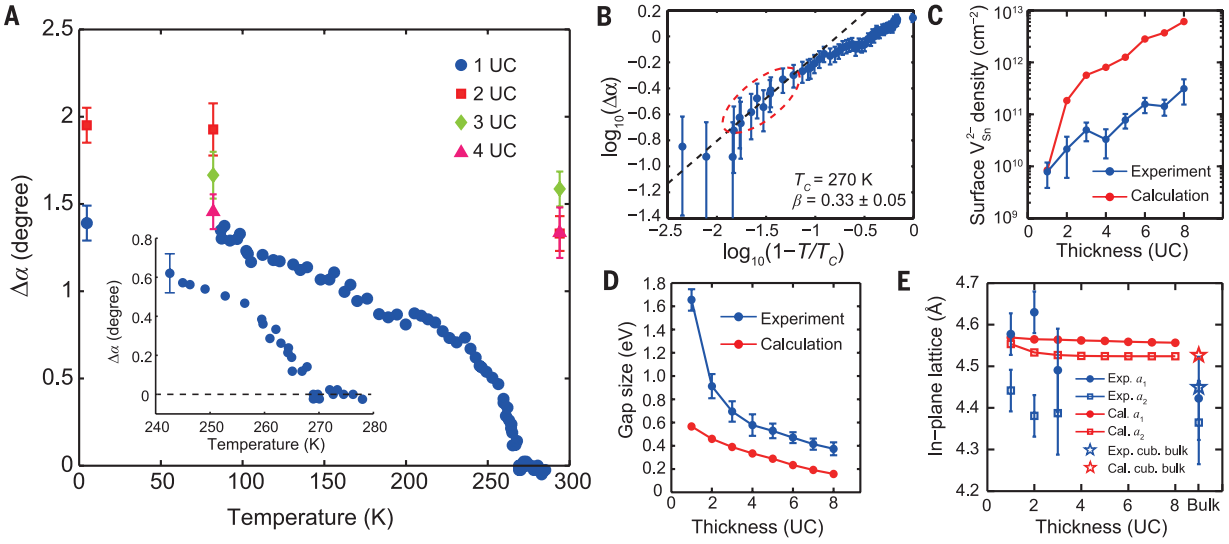



**Fig. 3. $T_c$ enhancement in the 1- to 4-UC SnTe films.** (**A**) Temperature dependence of the distortion angle for the 1- to 4-UC SnTe films. (Inset) The distortion angle near $T_c$ = 270 K for the 1-UC film exhibiting the behavior of a second-order phase transition. (**B**) Power-law fitting of the data for 1-UC SnTe film near $T_c$. The critical exponent of $P$ is derived using the approximation $P^2 \propto \Delta\alpha$. The red dashed-line ellipse indicates the data points for linear fitting. (**C**) Thickness dependence of Sn vacancy density at the growth temperature of 450 K. (**D**) Experimentally measured and DFT-calculated thickness dependence of the electronic band gap. The discrepancy between experimental data and DFT calculations comes from the well-known underestimation of band gap in the DFT method (see figs. S23 and S24). (**E**) Thickness dependence of the lattice constants determined by experimental data and DFT calculations (see also fig. S26).

minimize the electrostatic energy (*32*). Occasionally, the in-plane 180° domain wall is also observed (Fig. 2E). More images of domain structures can be found in figs. S14 and S15.

One more important criterion for ferroelectricity is that the polarization can be manipulated by electric field (*32*). A ferroelectric crystal should have two or more orientational states for polarization, which can be shifted from one to another by an electric field. To tune the ferroelectric state, we applied a 5-V voltage pulse between the STM tip and the 1-UC SnTe film. The domain structure of the film is distinctly changed by the pulse (compare the upper and lower panel of Fig. 2E). Here, STM manipulates the ferroelectric polarization through domain wall motion instead of rotation of polarization. The electric field in the tunneling junction is comparable to the typical activation field (~$10^7$ V/m) for moving a ferroelectric domain wall. Such tunability distinguishes ferroelectric from other polar states.

The above studies, including the formation of domain structure, lattice distortion, band-bending, and polarization manipulation by electric field, strongly support the occurrence of ferroelectricity in the 1-UC SnTe film. However, all the evidence so far is still at the liquid helium temperature. At higher temperatures, the spontaneous polarization $\vec{P}$ diminishes and eventually disappears at the ferroelectric transition temperature $T_c$, as does the domain structure (fig. S17). During the variable temperature measurement, the distortion angle $\Delta\alpha$ of the rock-salt unit cell is conveniently determined by the separation between the two sets of Bragg peaks (fig. S18). The polarization is then derived through the relation $P^2 \propto \Delta\alpha$ (*33*). The temperature dependence of $\Delta\alpha$ is plotted in Fig. 3A. For the 1-UC film, $\Delta\alpha$ drops as the temperature increases and becomes zero at $T_c$ = 270 K. The critical temperature for 1-UC film is much higher than the bulk value of about 100 K (*25*). Within the precision of measurement, $\Delta\alpha$ approaches zero continuously at $T_c$ (Fig. 3A, inset), which is in agreement with the behavior for a second-order phase transition. The critical index $\beta = 0.33 \pm 0.05$ [$P \sim (T_c - T)^{\beta}$] is extracted by plotting $\Delta\alpha$ versus $T$ on a log-log scale (Fig. 3B). The critical exponent here is identical to the value, 0.33, observed in the $PbZr_{0.9}Ti_{0.1}O_3$ bulk material (*34*) and larger than the values of currently available 2D models with short-range interaction, such as 1/8 of the 2D Ising model (*32*). It indicates that a long-range correlated microscopic model is required to fully account for our observations. From β alone, the universality class of 1-UC SnTe cannot be determined.

For thicker films from 2 to 4 UC, $T_c$ is even higher than the room temperature (RT). Variable-temperature STM measurement above RT is beyond the capability of our current instrument. However, we observed both domain structures and lattice distortion (fig. S19), as well as the band-bending (fig. S20) at RT for 2- to 4-UC films. The evidence for a ferroelectric phase at RT also comes from Raman spectroscopy. The ferroelectric phase transition can be viewed as a condensation of transverse optical (TO) phonon near the Brillouin zone center. Raman spectroscopy directly probes the TO mode softening in ferroelectrics. For SnTe, the Raman signal is inactive above $T_c$ because of the crystalline symmetry of the rock-salt structure. Below $T_c$, the TO mode becomes Raman active, and its frequency is given by $\omega_{TO} \propto (T_c - T)^{1/2}$ for bulk (*23*). The spectra of 2-UC SnTe film show that the TO mode persists up to RT with only slight softening (fig. S21). For comparison, $T_c$ of the 20-nm-thick film extracted from the TO mode peak shift (fig. S22) is found to be 130 K and consistent with a previous report (*23*).

The strong ferroelectricity enhancement in atomic-layer-thick SnTe films is unusual and may have its origin in the lower free carrier density. The ferroelectric transition strongly depends on the screening effect of free carriers on the dipole-dipole interaction [see (*28*) and fig. S25]. For example, the bulk transition temperature of SnTe reaches 100 K only when the carrier density has been reduced to $10^{20}$ $cm^{-3}$. The carrier density depends on the density of defects and the size of the band gap. Both experiment and density functional theory (DFT) calculation (Fig. 3C) reveal that the density of Sn vacancies drops by 2 to 3 orders of magnitude in the SnTe ultrathin film. Besides the lower density of defects, the larger energy gap also helps to make SnTe thin film less conductive. The bulk IV-VI semiconductors such as SnTe usually have narrow band gaps. However, the gap of SnTe is dramatically increased when the film thickness is less than 8 UC (Fig. 3D). The band gap enhancement partially comes from the quantum confinement in the ultrathin film. The $T_c$ of 1-UC film is lower than that of the 2- to 4-UC films, probably because the effects of reduced dimensionality become more prominent in the case of 1 UC.

The in-plane lattice expansion may also facilitate the $T_c$ enhancement. Both experiment (with slightly large uncertainty) and the DFT calculation suggest that the in-plane lattice constants of SnTe increase when the film becomes thinner (Fig. 3E), as generally occurs in the thin film with rock-salt structure (*35*). The lattice constants tune ferroelectricity by adjusting the balance between long-range Coulomb interaction and short-range repulsion (*36*). Such mechanisms may partially

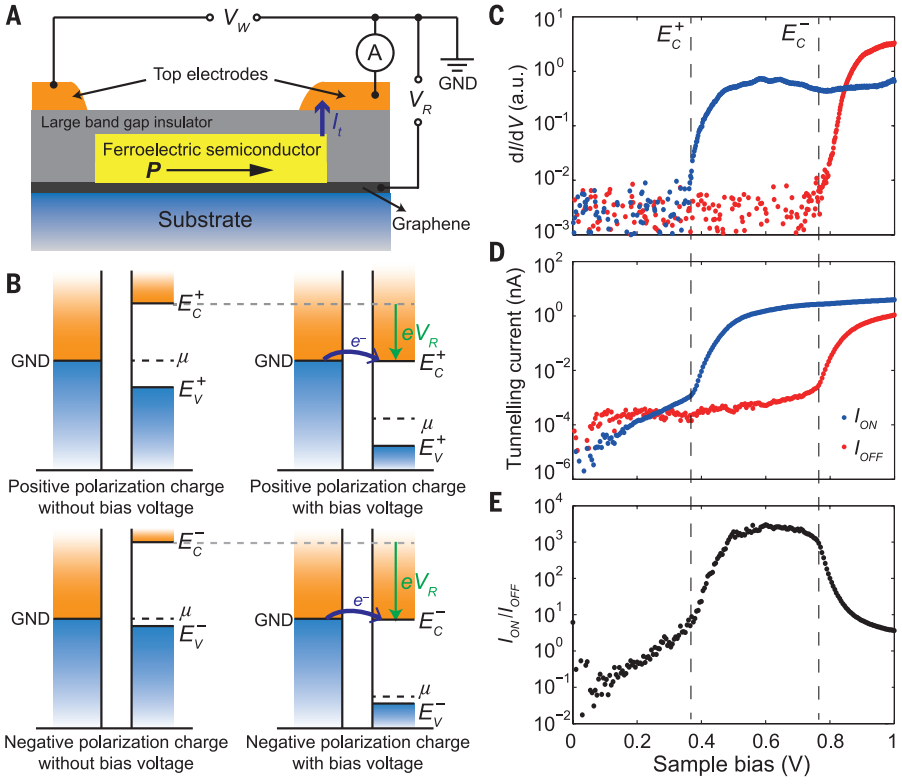



**Fig. 4. Proposal for a nonvolatile memory device based on the ferroelectric tunnel junction with in-plane polarization.** (**A**) Schematic of the device structure. $V_R$ and $V_W$ are the reading and writing voltages; $I_t$ is the tunneling current. (**B**) Band diagram of the tunneling process between the ferroelectric thin film and a top electrode. $\mu$ is the chemical potential of the film. For the positively (negatively) charged state, a threshold voltage $V_R = E_C^+/e(E_C^-/e)$ is needed to open the tunneling channel between the top electrode and the conduction band of the edge. (**C**) Threshold voltages measured by STM. For a 3-UC film, $E_C^+$ and $E_C^-$ are measured by d*I*/d*V* (set point: 1.0 V, 100 pA) and found to be 0.36 and 0.76 eV, respectively. (**D**) Simulation of the reading process by STM in a 3-UC film. When the tip moves from one edge to another, the feedback loop of STM is turned off to maintain the same distance of tunneling junction. A large on/off ratio is achieved between $E_C^+/e$ and $E_C^-/e$. (**E**) Bias dependence of the on/off ratio for a 3-UC film obtained from the data in Fig. 4D.

account for the enhancement of ferroelectricity in SnTe and perovskite thin films (*12*, *37*, *38*). In particular, SnTe thin films share several common features with $SrTiO_3/DyScO_3$ (*38*): in-plane lattice expansion, out-of-plane lattice contraction, and in-plane polarization.

Based on the in-plane polarized ferroelectric thin film, a nonvolatile ferroelectric random access memory (FeRAM) device (Fig. 4A) can be designed to take advantage of the small size and high transition temperature. The voltage pulses $V_W$ and $V_R$ are applied only during the writing and reading processes, respectively. The writing voltage $V_W$ generates the in-plane electric field to flip the in-plane polarization of the ferroelectric film. The two opposite directions of polarization represent the "on" and "off" states of the memory unit. The states are read by electron tunneling into an edge driven by the voltage $V_R$. The tunneling current $I_t$ strongly depends on the band-bending. The dependence can be easily understood by the band structure on the edges (Fig. 4, B and C). To demonstrate the mechanism and simulate the reading process, the *I*-*V* (current-voltage) curves (Fig. 4D) on the edges of a 3-UC SnTe film were measured by STM. During the measurement, the gap distance between tip and sample was fixed. The tunneling current increases rapidly after the bias voltage reaches the corresponding threshold $V_R$. From 0.5 to 0.7 eV, the on/off ratio can reach as high as 3000 (Fig. 4E). Similar measurement is also performed on 1-UC film, which shows a much lower on/off ratio. Compared with the conventional FeRAM, where reading is destructive, the memory based on the in-plane polarization and tunneling does not reverse the polarization and is nondestructive. Moreover, the fact that ferroelectricity and band-bending could exist in the SnTe nanowire of only 16-nm width (fig. S27) shows the potential to fabricate devices with high density. The materials for insulating layers and electrodes need to be carefully selected for achieving optimal performance.

## REFERENCES AND NOTES


1. K. S. Novoselov *et al.*, *Nature* **438**, 197–200 (2005).
2. Y. Zhang, Y.-W. Tan, H. L. Stormer, P. Kim, *Nature* **438**, 201–204 (2005).
3. A. Gozar *et al.*, *Nature* **455**, 782–785 (2008).
4. T. Zhang *et al.*, *Nat. Phys.* **6**, 104–108 (2010).
5. W. Dürr *et al.*, *Phys. Rev. Lett.* **62**, 206–209 (1989).
6. M. Farle, K. Baberschke, *Phys. Rev. Lett.* **58**, 511–514 (1987).
7. X. Xi *et al.*, *Nat. Nanotechnol.* **10**, 765–769 (2015).
8. M. M. Ugeda *et al.*, *Nat. Phys.* **12**, 92–97 (2016).
9. T. Tybell, C. H. Ahn, J.-M. Triscone, *Appl. Phys. Lett.* **75**, 856 (1999).
10. D. D. Fong *et al.*, *Science* **304**, 1650–1653 (2004).
11. D. D. Fong *et al.*, *Phys. Rev. Lett.* **96**, 127601 (2006).
12. D. A. Tenne *et al.*, *Phys. Rev. Lett.* **103**, 177601 (2009).
13. P. Ghosez, K. M. Rabe, *Appl. Phys. Lett.* **76**, 2767 (2000).
14. J. Junquera, P. Ghosez, *Nature* **422**, 506–509 (2003).
15. Z. Wu *et al.*, *Phys. Rev. B* **70**, 104108 (2004).
16. N. Sai, A. M. Kolpak, A. M. Rappe, *Phys. Rev. B* **72**, 020101 (2005).
17. N. Sai, C. J. Fennie, A. A. Demkov, *Phys. Rev. Lett.* **102**, 107601 (2009).
18. Y. Zhang, G.-P. Li, T. Shimada, J. Wang, T. Kitamura, *Phys. Rev. B* **90**, 184107 (2014).
19. E. Almahmoud, I. Kornev, L. Bellaiche, *Phys. Rev. B* **81**, 064105 (2010).
20. A. V. Bune *et al.*, *Nature* **391**, 874–877 (1998).
21. J. O. Dimmock, I. Melngailis, A. J. Strauss, *Phys. Rev. Lett.* **16**, 1193–1196 (1966).
22. N. Wang *et al.*, *Phys. Rev. B* **89**, 045142 (2014).
23. S. Sugai, K. Murase, H. Kawamura, *Solid State Commun.* **23**, 127–129 (1977).
24. W. Jantsch, *Dielectric Properties and Soft Modes in Semiconducting (Pb, Sn, Ge) Te* (Springer, 1983).
25. K. L. I. Kobayashi, Y. Kato, Y. Katayama, K. F. Komatsubara, *Phys. Rev. Lett.* **37**, 772–774 (1976).
26. M. Iizumi, Y. Hamaguchi, K. F. Komatsubara, Y. Kato, *J. Phys. Soc. Jpn.* **38**, 443–449 (1975).
27. T. H. Hsieh *et al.*, *Nat. Commun.* **3**, 982 (2012).
28. See supplementary materials on *Science* Online.
29. Z. Sheng, N. Ogawa, Y. Ogimoto, K. Miyano, *Adv. Mater.* **22**, 5507–5511 (2010).
30. V. Garcia *et al.*, *Nature* **460**, 81–84 (2009).
31. J. L. Giocondi, G. S. Rohrer, *Top. Catal.* **49**, 18–23 (2008).
32. M. E. Lines, A. M. Glass, *Principles and Applications of Ferroelectrics and Related Materials* (Oxford Univ. Press, 1977).
33. R. Clarke, *Phys. Rev. B* **18**, 4920–4926 (1978).
34. R. Clarke, A. M. Glazer, *J. Phys. C Solid State Phys.* **7**, 2147–2156 (1974).
35. F. W. de Wette, W. Kress, U. Schröder, *Phys. Rev. B* **32**, 4143–4157 (1985).
36. G. A. Samara, T. Sakudo, K. Yoshimitsu, *Phys. Rev. Lett.* **35**, 1767–1769 (1975).
37. K. J. Choi *et al.*, *Science* **306**, 1005–1009 (2004).
38. J. H. Haeni *et al.*, *Nature* **430**, 758–761 (2004).


## ACKNOWLEDGMENTS

We thank H. J. Xiang, Y. G. Zhao, P. Yu, G. M. Zhang, and B. F. Zhu for helpful discussions. We thank Y. Wang for sample characterization by transition electron microscope and K. Li for helping to prepare substrates. The work is financially supported by National Natural Science Foundation and Ministry of Science and Technology of China. Q.Z. was supported by the Ministry of Science and Technology of China (973 projects 2012CB921701) and the NSF of China. J.L. was supported by the STC Center for Integrated Quantum Materials, NSF Grant No. DMR-1231319. L.F. was supported by the DOE Office of Basic Energy Sciences, Division of Materials Sciences and Engineering under Award No. DE-SC0010526. A provisional patent based on the concept and device in this work has been filed at MIT.

## SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/353/6296/274/suppl/DC1
Materials and Methods
Supplementary text
Figs. S1 to S27
Table S1
References (*39*–*51*)

13 November 2015; accepted 16 June 2016
10.1126/science.aad8609

ORGANIC CHEMISTRY

# Aryl amination using ligand-free Ni(II) salts and photoredox catalysis

Emily B. Corcoran,[1] Michael T. Pirnot,[2] Shishi Lin,[3] Spencer D. Dreher,[3] Daniel A. DiRocco,[3] Ian W. Davies,[3] Stephen L. Buchwald,[2]* David W. C. MacMillan[1]*

**Over the past two decades, there have been major developments in transition metal–catalyzed aminations of aryl halides to form anilines, a common structure found in drug agents, natural product isolates, and fine chemicals. Many of these approaches have enabled highly efficient and selective coupling through the design of specialized ligands, which facilitate reductive elimination from a destabilized metal center. We postulated that a general and complementary method for carbon–nitrogen bond formation could be developed through the destabilization of a metal amido complex via photoredox catalysis, thus providing an alternative approach to the use of structurally complex ligand systems. Here, we report the development of a distinct mechanistic paradigm for aryl amination using ligand-free nickel(II) salts, in which facile reductive elimination from the nickel metal center is induced via a photoredox-catalyzed electron-transfer event.**

Transition metal–catalyzed cross-coupling reactions are fundamental methods for constructing complex molecular architectures. Key among these reactions has been the formation of C–N bonds to access anilines, a common structure found in medicinal agents and other complex molecules, through the coupling of amines with aryl halides or pseudohalides (*1*–*6*). Over the past two decades, multiple generations of ligands have been designed that have elevated palladium-catalyzed aryl amination to an essential transformation with both predictable reactivity and practical importance. The ligand frameworks that have been developed for palladium catalysis often employ a few key design elements (i.e., sterically encumbered architecture, hemilabile coordinating groups, electronically tuned substituents) to enact high catalytic efficiency (*7*). One essential feature in ligand design is the capacity for destabilization of the Pd(II) amido complex, which induces reductive elimination resulting in C–N bond formation (*8*–*10*).

The broad success of ligated-palladium catalysis has hinged on the ability to selectively design, synthesize, and implement ligand classes to enable specific coupling reactions in high levels of efficiency and selectivity. Inspired by the effectiveness of structurally diverse ligand systems on palladium reductive elimination, we recognized that a general and complementary approach for productive C–N bond formation could be developed through an alternative mechanistic approach. In particular, we envisioned that the destabilization of a metal amido species could occur through an electron transfer pathway via photoredox catalysis, thus inducing facile C–N bond formation analogous to the role of ligand in traditional palladium couplings (Fig. 1).

We proposed that nickel could be an ideal transition metal for this catalytic platform because of its known proclivity to participate in one-electron processes (*11*–*13*). Furthermore, this ligand-free mechanism could potentially broaden the applicability of nickel catalysis in aryl amination (*14*–*18*), a system that is often plagued with the use of harsh reductants and air-sensitive Ni(0) complexes. Nickel(II) amido complexes, as detailed by Hillhouse and others, tend not to undergo reductive elimination at ambient temperature (*19*–*21*). Computations suggest that C–N reductive elimination from Ni(II) alkyl amido complexes is thermodynamically disfavored, in contrast to the exothermic reductive elimination from analogous Pd(II) alkyl amido complexes (*22*). Exposure of Ni(II) amido complexes to various oxidants results in facile reductive elimination, presumably through single-electron oxidation of the Ni(II) amido complex to the respective Ni(III) complex, lending support to our design strategy (*19*, *20*).

In recent years, the merging of photoredox catalysis with nickel catalysis has enabled both C–O (*23*) and C–S (*24*, *25*) bond formation through the generation of reactive radical intermediates and/or the single-electron modulation of nickel oxidation states (*26*, *27*). The feasibility of these reactions arises from the capacity of photocatalysts to act as both strong single-electron oxidants and reductants upon irradiation with visible light (*28*). Critical to the success of our design plan, the photoredox catalyst would be involved in two key aspects of our postulated mechanism: (i) reduction of a Ni(II) salt to the active Ni(0) catalyst through the use of a mild sacrificial reductant and (ii) electron transfer to destabilize the Ni(II) amido complex toward reductive elimination in the absence of an exogenous ligand.

In our proposed mechanism (Fig. 2), irradiation of heteroleptic Ir(III) photocatalyst Ir[dF(CF)$_3$ppy]$_2$(dtbbpy)PF$_6$ [dF(CF$_3$)ppy = 2-(2,4-difluorophenyl)-5-(trifluoromethyl)pyridine; dtbbpy = 4,4′-di-*tert*-butyl-2,2′-bipyridine] (**1**) with visible light would produce long-lived triplet photoexcited state *Ir$^{III}$ **2** ($\tau$ = 2.3 μs) (*29*). Concomitantly, Ni(0) complex **3** would oxidatively insert into bromoarene **4** to yield the corresponding Ni(II) aryl bromide complex **5**. The Ni(II) complex would then undergo ligand exchange and subsequent deprotonation to arrive at Ni(II)-aryl amido complex **6**. This Ni(II) complex could then participate in a single-electron transfer (SET) event with the photoexcited *Ir$^{III}$ catalyst ($E_{1/2}^{red}$ [*Ir$^{III}$/Ir$^{II}$] = + 1.21 V versus saturated calomel electrode (SCE) in CH$_3$CN) to afford Ni(III) complex **7** and the reduced Ir(II) photocatalyst (**8**) (*29*). Reductive elimination of the resultant Ni(III) complex would yield Ni(I) bromide species **9** and desired aniline product **10**. Both the photoredox and nickel catalytic cycles are closed upon a SET event between Ni(I)Br species **9** and Ir(II) ($E_{1/2}^{red}$ [Ir$^{III}$/Ir$^{II}$] = −1.37 V versus SCE in CH$_3$CN) in which Ni(0) catalyst **3** and ground state Ir(III) catalyst **1** are regenerated simultaneously (*29*). The homogeneous nickel species is presumably coordinated to solvent, base, and/or amine throughout the process (denoted as L$_n$ in Fig. 2), but would not require a designed ligand for efficient reactivity.

With this mechanistic hypothesis in hand, we first examined the proposed cross-coupling reaction with 4-bromobenzotrifluoride, pyrrolidine, Ni(II) bromide glyme, and a variety of ligands, bases, and solvents (see supplementary materials for details). The employment of photocatalyst **1**, 1,4-diazobicyclo[2.2.2] octane (DABCO) as a base and di-*tert*-butyl-2,2′-bipyridine as a supporting ligand provided the respective aniline product **10** in 72% yield in the presence of blue light. Control experiments established the importance of both nickel and photocatalyst, as no formation of the desired cross-coupled product was observed in the absence of either nickel, photocatalyst, or light. We sought to investigate the effectiveness of this protocol in the absence of ligand to confirm our initial mechanistic hypothesis. The cross-coupling reaction proceeded with enhanced reactivity in the absence of di-*tert*-butyl-2,2′-bipyridine, delivering the desired aniline product in 96% yield. This result supports our proposed mechanistic pathway in Fig. 2 and the capacity of visible light and a photocatalyst to initiate reductive elimination in the absence of ligand.

With the optimal conditions in hand, we sought to demonstrate the generality of this ligand-free amination. As highlighted in Fig. 3, a wide variety of functional groups were tolerated on the amine coupling partner, including protected nitrogen atoms, trifluoromethyl groups, alcohols, alkenes, and sulfonamides (**11** to **15**, 72 to 77% yield). The high efficiency of the photoredox catalytic cycle is evidenced by the low photocatalyst loadings [≤0.02 mole percent (mol %)] required to effect the desired cross-coupling reactions. Monofluorinated amines (**16**) coupled cleanly under these conditions in 81% yield. Coupling of amine

[1]Merck Center for Catalysis at Princeton University, Princeton, NJ 08544, USA. [2]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [3]Department of Process Chemistry, Merck Research Laboratories, Rahway, NJ 07065, USA.
*Corresponding author. Email: dmacmill@princeton.edu (D.W.C.M); sbuchwal@mit.edu (S.L.B.)

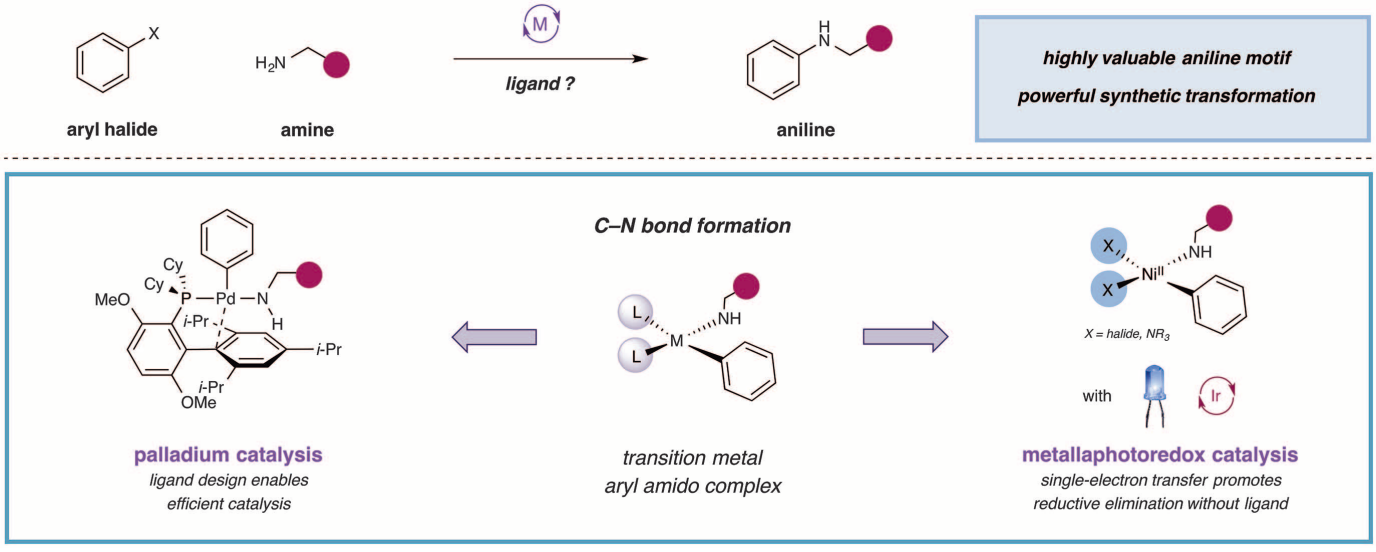



**Fig. 1. Proposed approach to C–N cross-coupling—promotion of reductive elimination via single-electron transfer.**

hydrogen chloride salts was accomplished in good yield by using MTBD (7-methyl-1,5,7-diazabicyclo-[4.4.0]deca-5-ene) as the base (**16** and **17**, 81 and 70% yield); although the expense of MTBD often limits its practicality, synthetically useful yields were also obtained with DABCO in most cases (see supplementary materials for details). Chemoselective *N*-arylation in the presence of alcohols was also achieved, as no C–O coupling was observed (**13**, 76% yield, single regioisomer).

We found that α-substitution on the amine partner was tolerated in moderate yields (**20** and **22**, 60 and 78% yield, respectively), but hindered amines, e.g., *tert*-butylamine, did not productively couple. Although nucleophilic amines, such as pyrrolidine and morpholine, coupled in good efficiencies (**10** and **22**, 96 and 91% yield, respectively) at ambient temperature after only a few hours, less nucleophilic substrates, such as trifluoroethylamine (**12**, 77% yield), allylamine (**14**, 76% yield), and furfurylamine (**23**, 90% yield), required longer reaction times (>24 hours). Amines lacking α-hydrogens failed to react under these reaction conditions. We hypothesized that an initial β-hydride elimination event is a prerequisite for accessing the active Ni(0) catalyst in this reaction. Indeed, the addition of a substoichiometric amount of amine possessing α-hydrogens, such as pyrrolidine, to the reaction resulted in good yields (72 and 84% yield) of cross-coupled products **15** and **24**, respectively, with less than 10% of the pyrrolidine-coupled product being observed. For a number of substrates, product formation was observed in high yield with $Ru(bpy)_3(PF_6)_2$ as the photocatalyst (see fig. S8). In cases where the yield is lower, protodehalogenation, along with trace phenol, aryl ether, or aryl chloride formation, can account for the mass balance of the reaction.

A variety of aryl bromides containing diverse functional groups, such as nitriles, amides, trifluoromethyl groups, halides, esters, and sulfonamides, performed well under the reaction conditions (**25** to **31**, 70 to 93% yield), with substitution at the ortho position being tolerated (**26** and **32**, 83 and 73% yield, respectively) (Fig. 3). Six-membered heteroaromatic substrates were effectively coupled in good yields (**27** and **33** to **35**, 71 to 91% yield). Unfortunately, a series of aryl bromides, e.g., select heterocycles and a 2,6-disubstituted arene, failed to couple under the current reaction conditions (see fig. S8). For electron-rich arenes (**32** and **36**), lowering the photocatalyst loading (0.002 mol %) increased reaction efficiency and inhibited competitive protodehalogenation, as did the use of MTBD as base. Additionally, the initial evaluation of this C–N

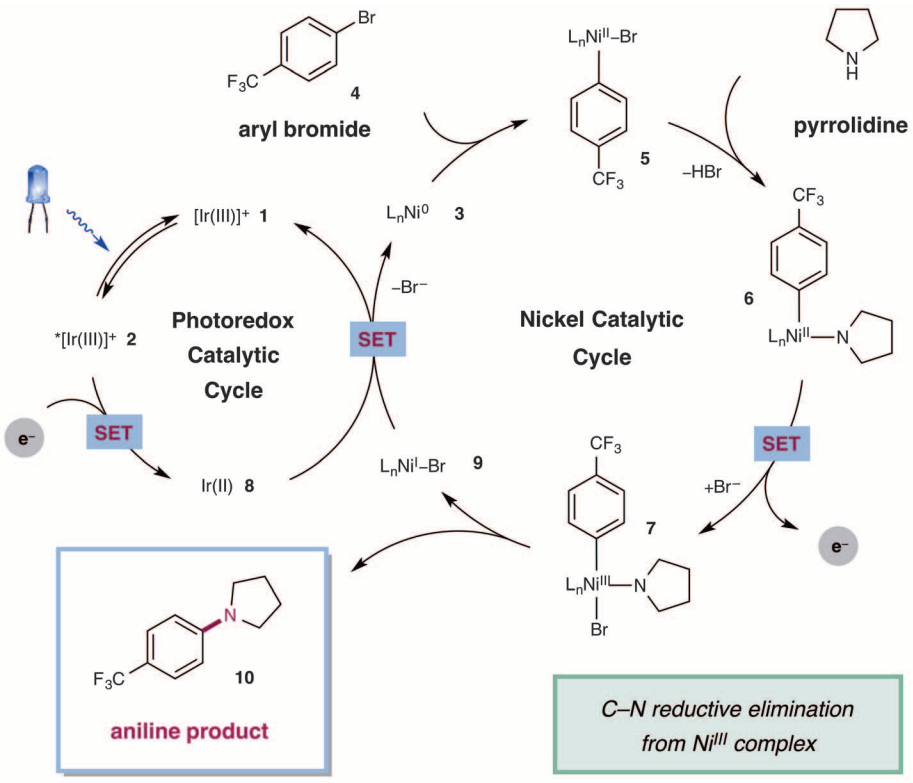



**Fig. 2. Proposed mechanism for the metallaphotoredox-catalyzed amination reaction.** L, DABCO or amine; SET, single-electron transfer.

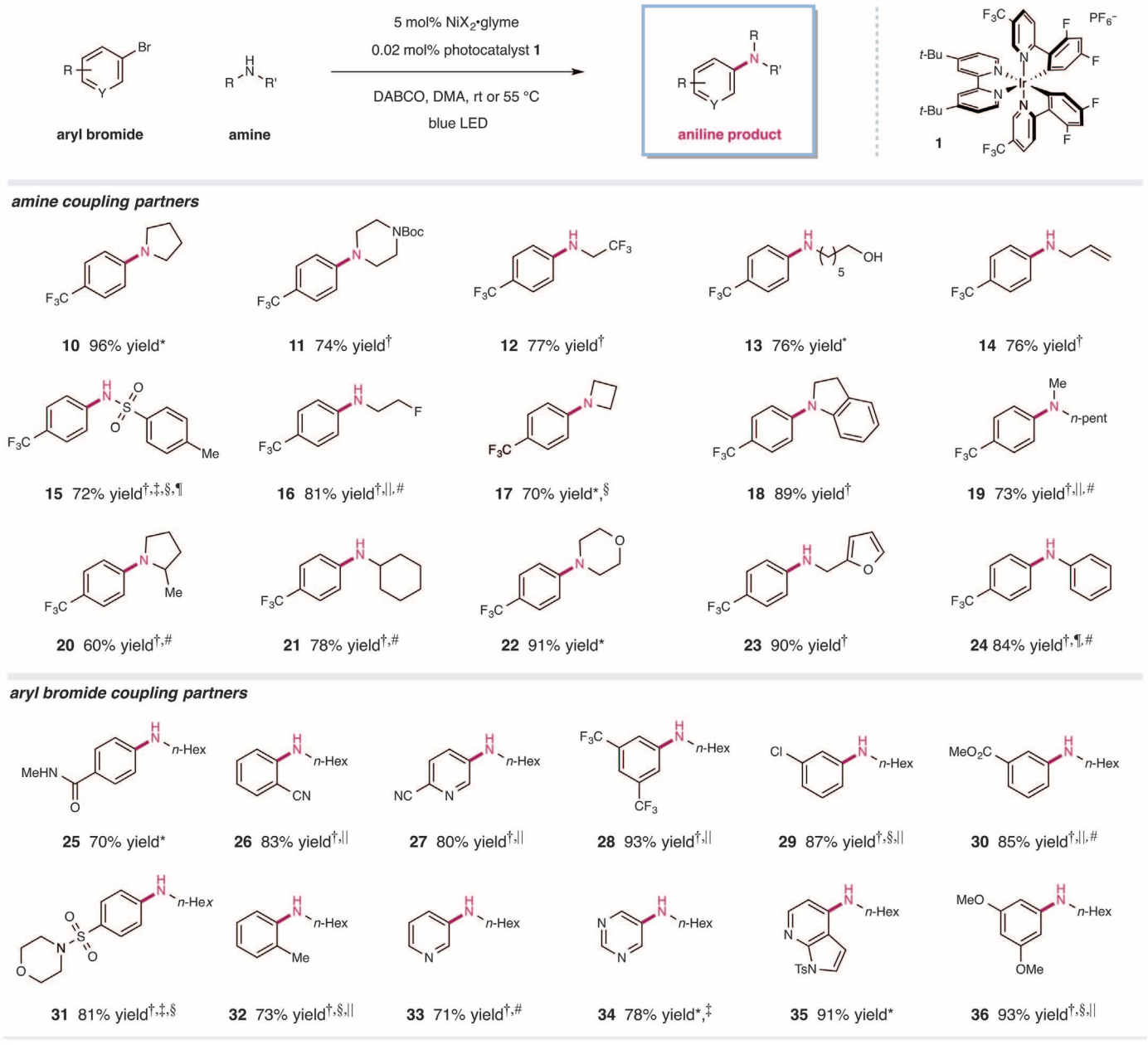



**Fig. 3. Metallaphotoredox-catalyzed amination: amine and arene scope.** For each entry number (in bold), data are reported as percent isolated yield. R: H, alkyl, or aryl substrate; Y: C, CH, or N; X: Cl or Br; DABCO: 1,4-diazabicyclo[2.2.2]octane; DMA, *N*,*N*-dimethylacetamide; rt, room temperature; LED: light-emitting diode; Me: methyl; *n*-pent: *n*-pentyl; Boc: *tert*-butoxycarbonyl. *Run at ambient temperature. †Reaction heated to 55°C. ‡DMSO (dimethyl sulfoxide) used as solvent. §MTBD (7-methyl-1,5,7-triazabicyclo[4.4.0]dec-5-ene) used as base. ‖0.002 mol % **1** used. ¶10 mol % pyrrolidine included. #See supplementary materials for details.

cross-coupling platform in flow chemistry has shown high efficiency in decreased reaction times (see supplementary materials).

To investigate the role of the photocatalyst in switching on nickel-catalyzed aryl amination, we used $Ni(cod)_2$. In the absence of light, the formation of aniline **10** was severely diminished (10% yield), whereas standard reaction conditions yielded **10** in 91% yield. This result supports our hypothesis that the cross-coupling is light-mediated and the effect of the photoredox catalyst is not limited to the reduction of a Ni(II) salt to the active Ni(0) species.

Lastly, this C–N cross-coupling platform was evaluated in a high-throughput fashion with a microscale informer plate as developed by Merck laboratories. A chemistry informer library of 18 complex, druglike aryl halides was subjected to Ni-photoredox–based amination conditions with piperidine. Notably, 78% of substrates underwent successful cross-coupling and provided "reaction hits" that were suitable for optimization (Fig. 4) (*30*). Benchmarked against other protocols that have been evaluated with this informer series, this study represents one of the most successful (*31*) generic catalysis platforms for aryl amination that has been evaluated within Merck. On the basis of these results, we are confident that this complementary method will find broad applicability in the synthesis of structurally diverse, druglike compounds.

Through the merger of photoredox and nickel catalysis, challenging C–N reductive elimination can be switched on through the use of visible light and a photocatalyst via the intermediacy of a Ni(III) oxidation state (*32*–*34*). This strategy represents a complementary approach to traditional ligated-palladium catalysis through the use of a distinct mechanistic pathway for reductive

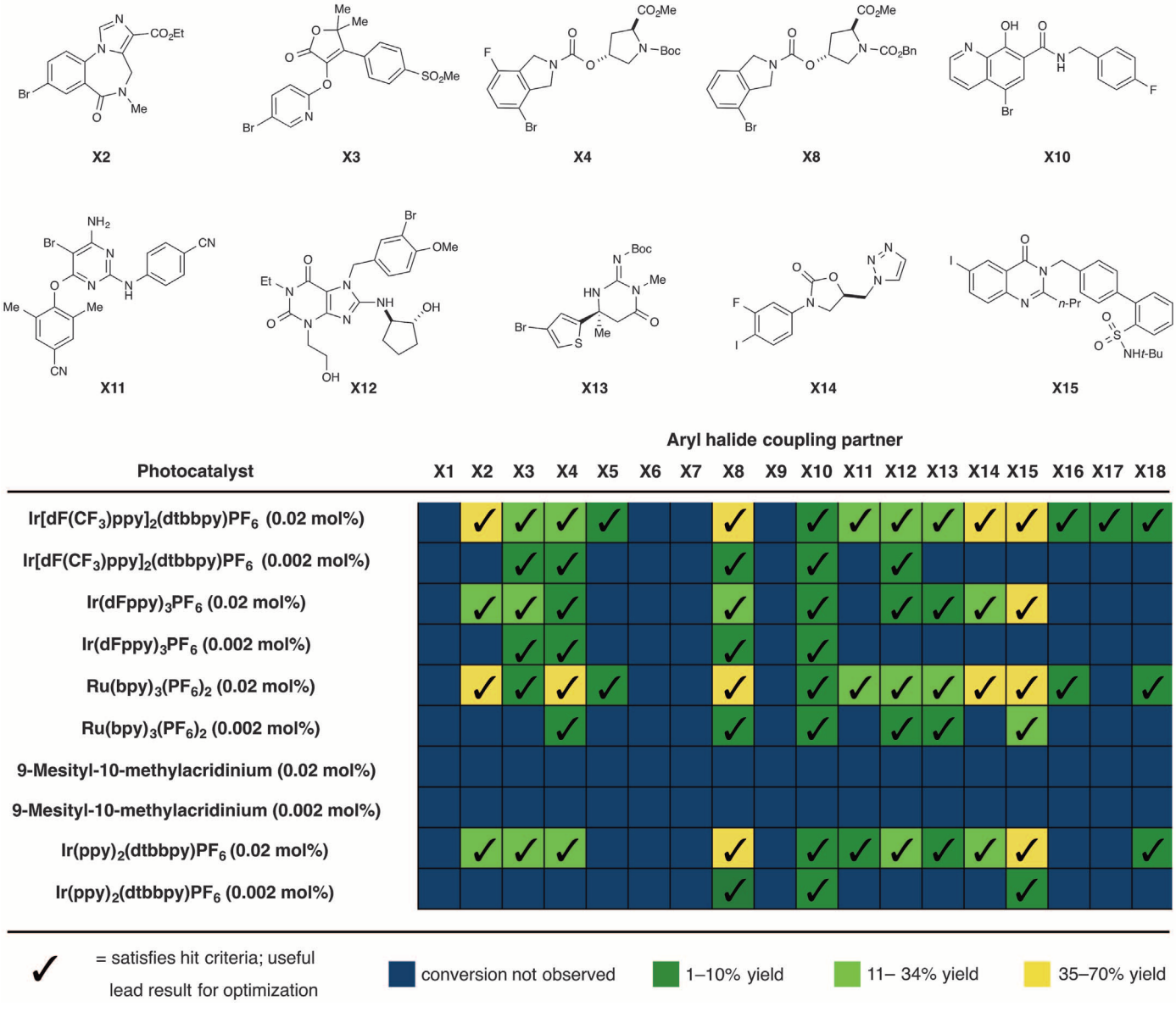



**Fig. 4. High-throughput metallaphotoredox-catalyzed aryl amination with a chemistry informer library.** For each compound numbered in bold, the cross-coupling was performed with piperidine using DABCO as a base. Ac, acetyl; Bn, benzyl; Boc, *tert*-butoxycarbonyl; Et, ethyl; Me, methyl; *n*-Pr, *n*-propyl; *t*-Bu, *tert*-butyl. Conversion was quantified by ultrapressure liquid chromatography/mass spectrometry (UPLC/MS) using an internal standard. See supplementary materials for exact yields and conditions.

elimination, which will likely be broadly applicable across a range of substrate classes.

## REFERENCES AND NOTES


1. F. Ullmann, *Chem. Ber.* **36**, 2382–2384 (1903).
2. I. Goldberg, *Chem. Ber.* **39**, 1691–1692 (1906).
3. M. Kosugi, M. Kameyama, T. Migita, *Chem. Lett.* **12**, 927–928 (1983).
4. M. S. Driver, J. F. Hartwig, *J. Am. Chem. Soc.* **118**, 7217–7218 (1996).
5. D. Chan, *Tetrahedron Lett.* **37**, 9013–9016 (1996).
6. A. S. Guram, S. L. Buchwald, *J. Am. Chem. Soc.* **116**, 7901–7902 (1994).
7. D. S. Surry, S. L. Buchwald, *Chem. Sci.* **2**, 27–50 (2011).
8. B. P. Fors, D. A. Watson, M. R. Biscoe, S. L. Buchwald, *J. Am. Chem. Soc.* **130**, 13552–13554 (2008).
9. J. F. Hartwig, *Synlett* **1997**, 329–340 (1997).
10. A. T. Brusoe, J. F. Hartwig, *J. Am. Chem. Soc.* **137**, 8460–8468 (2015).
11. T. T. Tsou, J. K. Kochi, *J. Am. Chem. Soc.* **101**, 6319–6332 (1979).
12. The use of ligated nickel and photoredox catalysis for C–N bond formation has been reported (*13*).
13. S. Z. Tasker, T. F. Jamison, *J. Am. Chem. Soc.* **137**, 9531–9534 (2015).
14. J. P. Wolfe, S. L. Buchwald, *J. Am. Chem. Soc.* **119**, 6054–6058 (1997).
15. C.-Y. Gao, L.-M. Yang, *J. Org. Chem.* **73**, 1624–1627 (2008).
16. N. H. Park, G. Teverovskiy, S. L. Buchwald, *Org. Lett.* **16**, 220–223 (2014).
17. S. Ge, R. A. Green, J. F. Hartwig, *J. Am. Chem. Soc.* **136**, 1617–1627 (2014).
18. T. Shimasahi, M. Tobisu, N. Chatani, *Angew. Chem. Int. Ed.* **49**, 2929 (2010).
19. K. Koo, G. L. Hillhouse, *Organometallics* **14**, 4421–4423 (1995).
20. B. L. Lin, C. R. Clough, G. L. Hillhouse, *J. Am. Chem. Soc.* **124**, 2890–2891 (2002).
21. L. Ilies, T. Matsubara, E. Nakamura, *Org. Lett.* **14**, 5570–5573 (2012).
22. S. A. Macgregor, G. W. Neave, C. Smith, *Faraday Discuss.* **124**, 111–127, discussion 145–153, 453–455 (2003).
23. J. A. Terrett, J. D. Cuthbertson, V. W. Shurtleff, D. W. C. MacMillan, *Nature* **524**, 330–334 (2015).
24. M. S. Oderinde, M. Frenette, D. W. Robbins, B. Aquila, J. W. Johannes, *J. Am. Chem. Soc.* **138**, 1760–1763 (2016).
25. M. Jouffroy, C. B. Kelly, G. A. Molander, *Org. Lett.* **18**, 876–879 (2016).
26. J. C. Tellis, D. N. Primer, G. A. Molander, *Science* **345**, 433–436 (2014).
27. Z. Zuo *et al.*, *Science* **345**, 437–440 (2014).
28. C. K. Prier, D. A. Rankic, D. W. C. MacMillan, *Chem. Rev.* **113**, 5322–5363 (2013).
29. M. S. Lowry *et al.*, *Chem. Mater.* **17**, 5712–5719 (2005).

30. Previous studies leveraging high-throughput experimentation at Merck validated 5% yield as a good screening hit and 20% yield as a robust hit.
31. P. S. Kutchukian *et al*., *Chem. Sci. (Camb.)* **7**, 2604–2613 (2016).
32. $NiCl_2$ has been reported as a heterogeneous catalyst for the *N*-arylation of aryl iodides under microwave conditions in the absence of an exogenous ligand (*33*).
33. A. K. Gupta, G. T. Rao, K. N. Singh, *Tetrahedron Lett.* **53**, 2218–2221 (2012).
34. Although we favor the mechanism outlined in Fig. 2, we cannot rule out the possibility of energy transfer or direct excitation of Ni(II) in the presence of visible light.

**ACKNOWLEDGMENTS**

Research reported in this publication was supported by the NIH under award numbers GM58160 and RO1-GM078201-05. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. M.T.P. thanks the NIH for a postdoctoral fellowship (GM113311).

**SUPPLEMENTARY MATERIALS**

www.sciencemag.org/content/353/6296/279/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S12
References (*35*–*48*)
NMR Spectra

30 April 2016; accepted 3 June 2016
Published online 23 June 2016
10.1126/science.aag0209

GLACIERS

# Ocean forcing of glacier retreat in the western Antarctic Peninsula

**A. J. Cook,[1,2]* P. R. Holland,[3] M. P. Meredith,[3] T. Murray,[1] A. Luckman,[1] D. G. Vaughan[3]**

**In recent decades, hundreds of glaciers draining the Antarctic Peninsula (63° to 70°S) have undergone systematic and progressive change. These changes are widely attributed to rapid increases in regional surface air temperature, but it is now clear that this cannot be the sole driver. Here, we identify a strong correspondence between mid-depth ocean temperatures and glacier-front changes along the ~1000-kilometer western coastline. In the south, glaciers that terminate in warm Circumpolar Deep Water have undergone considerable retreat, whereas those in the far northwest, which terminate in cooler waters, have not. Furthermore, a mid-ocean warming since the 1990s in the south is coincident with widespread acceleration of glacier retreat. We conclude that changes in ocean-induced melting are the primary cause of retreat for glaciers in this region.**

The Antarctic Peninsula (AP) glaciers north of 70°S have the potential to raise sea level by 69 ± 5 mm (*1*), so any imbalance in their mass budget is of global importance. The region has undergone rapid warming in the latter half of the 20th century (*2*), and it is widely accepted that this has had a substantial impact on the ice sheet (*3*–*6*). The established theory that retreat of floating ice shelves is linked to a southerly migration of an atmospheric thermal limit (*7*) might also be considered likely to apply to retreat of marine-terminating glacier fronts. Indeed, the most significant glacier area loss over the past few decades has occurred in the northeast (*8*), which is north of the thermal limit and where the atmospheric temperature rise has been greatest.

Glaciers flowing westward from the AP plateau have, however, shown notable differences in frontal change over the same period. Overall ice loss has been greater in the south than the north, and glaciers in the northwest have remained stable (*8*). A previous study suggested that atmospheric warming may not be responsible for glacier change in this region because the migration from advance to retreat implied a warming more rapid than that observed (*9*). Indeed, spatial and temporal patterns of atmospheric forcings, including surface temperatures (*10*), melt duration (*4*), and precipitation (*11*), exhibit no clear relationship with the distinct north-south gradient of glacier-front changes along the west coast (*8*).

In the southwestern Bellingshausen Sea, rapid thinning of the ice shelves and their tributary glaciers has occurred during the past decade (*12*, *13*), and it has been proposed that this is caused by changes in upper-ocean heat content (*14*, *15*). The much larger ice loss from the West Antarctic Ice Sheet has also been linked to changes in heat content in the adjacent Amundsen Sea (*16*, *17*), which is similarly dominated by Circumpolar Deep Water (CDW) in its deeper layers. There, basal melting causes ice-shelf thinning, grounding-line retreat, and a loss of buttressing to the grounded ice inland. Variations in tidewater glacier termini are more complex, but several recent studies of Arctic glaciers have concluded that calving rates are strongly dependent on ocean temperatures [e.g., (*18*)]. Until now, the role of the ocean (as opposed to the atmosphere) as the dominant driver of glacier frontal retreat on the western AP has not been considered.

Although the oceans around Antarctica are notoriously data-sparse, the World Ocean Database 2013 (*19*) contains a sufficiently high spatial density of ocean temperature and salinity measurements to the west of the AP to enable regional mean temperature estimations (1945 to 2009) (*20*). When considered alongside observed changes in the glacier fronts (Fig. 1), a strong spatial correlation between the distribution of retreating glaciers and the pattern of mean ocean temperature over this period is revealed. Nearly all the glaciers south of Brabant and Anvers islands (~65°S), which discharge into warm ocean regions dominated by CDW, have suffered retreat. In contrast, the more northerly glaciers, which discharge into the cooler Bransfield Strait Water (BSW), experienced only small frontal changes indicative of relative stability over the 65 years for which observations are available. Furthermore, a southward increase in ice loss per glacier revealed in an earlier study (*8*) corresponds to a distinct and coherent spatial distribution in ocean temperatures (Fig. 2A). Ocean temperatures in this region are highly variable in the upper 100 m, but a pronounced north-south gradient becomes progressively more apparent at greater depth.

Partitioning the ocean adjacent to the AP into six regions of approximately equal area reveals three distinct oceanographic regimes (*21*, *22*) (Fig. 2, B and C). To the south and west, warm and saline CDW is prevalent across the Bellingshausen Sea shelf. This CDW is overlain by colder and fresher Winter Water (WW) and Antarctic Surface Water (AASW) formed by the interaction of CDW with the cryosphere and atmosphere. To the northeast, the Weddell Sea shelf contains cold and saline Shelf Water, which is heavily influenced by heat loss to the atmosphere and sea-ice production in the Weddell Sea. In the Bransfield Strait, northwest of the AP, the BSW is a mixture of Shelf Water and variants of CDW, again modified by air-sea-ice interaction. Crucially, these three water masses present very different thermal forcing to the glaciers abutting the ocean: the Shelf Water, BSW, and CDW average approximately 1°, 2°, and 4°C above the seawater freezing temperature, respectively (Fig. 2C). Glacier melting is expected to increase linearly or above-linearly with temperature above freezing, depending upon the geometry of the ice face and the presence of subglacial meltwater discharge (*23*, *24*).

The relationship between the ocean temperatures and glacier front change is also quantitatively robust (Fig. 3). Glaciers that have the warmest ocean temperatures near their fronts have retreated most significantly, and glaciers that are adjacent to the coolest water have remained stable or advanced. The relationship is strongly depth-dependent: Temperatures at and below 150-m depth display similar correlations,

[1]Department of Geography, Swansea University, Swansea, UK. [2]Department of Geography, Durham University, Durham, UK. [3]British Antarctic Survey, High Cross, Madingley Road, Cambridge, UK.
*Corresponding author. Email: alison.cook@durham.ac.uk

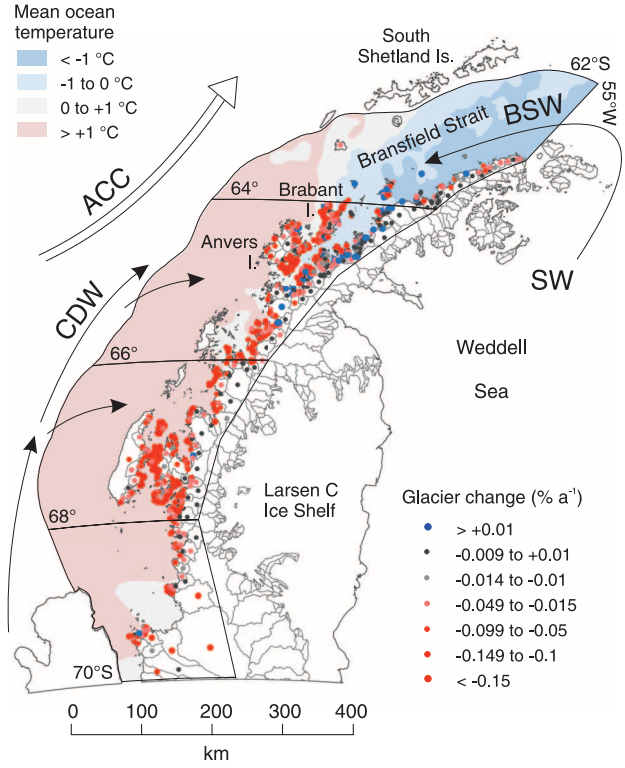



**Fig. 1. Mean ocean temperatures and overall glacier area changes, 1945 to 2009.** Mean in situ ocean temperature at 150-m depth (shaded) and glacier change (points). For each of the 674 glaciers along the west coast, the point shows overall change between its earliest and latest recorded ice-front position, relative to basin size (% $a^{-1}$). A similar spatial pattern is found for changes in absolute area loss per glacier. The point symbols are layered in the same order as in the legend (i.e., blue above red). Ocean circulation and water masses are also shown schematically: CDW, Shelf Water (SW), BSW, and ACC.

whereas at shallower depths there is no systematic relationship. The skewed nature of the glacier change rates (attributed to the wide range of glacier basin areas and characteristics) precludes linear regression, but when tested by rank order, strong correlations become apparent (table S1). Spearman's rank correlation between relative change rates (in % $a^{-1}$) and mean temperatures becomes stronger with depth. There is no correlation at 50 and 100 m, but deeper than this, the correlations are statistically significant ($P < 0.01$).

A relationship between ice-front history and deep ocean temperature is consistent with the expected dynamics of ocean melting of glacier ice in locations where the seabed is sufficiently deep. Several studies have shown that release of fresh, buoyant meltwater causes an upwelling at the ice face that draws in water at depth and drives a flow away from the glacier at the surface or pycnocline (*25*–*27*). Where present, this circulation preferentially delivers source waters for melting at depth, with melting comparatively insensitive to the properties of the shallower waters. The available bathymetric data for the western AP region support a connection between the deeper shelf waters and the glacier fronts (fig. S1).

We hypothesize that the spatial relationship between ocean properties and glacier change in the western AP is a consequence of the differing thermal characteristics of the two oceanographic regimes. The regimes have different temporal variabilities, with CDW variations originating in the transport and mixing of water from the Antarctic Circumpolar Current (ACC) and BSW variations originating in atmosphere-ocean interaction

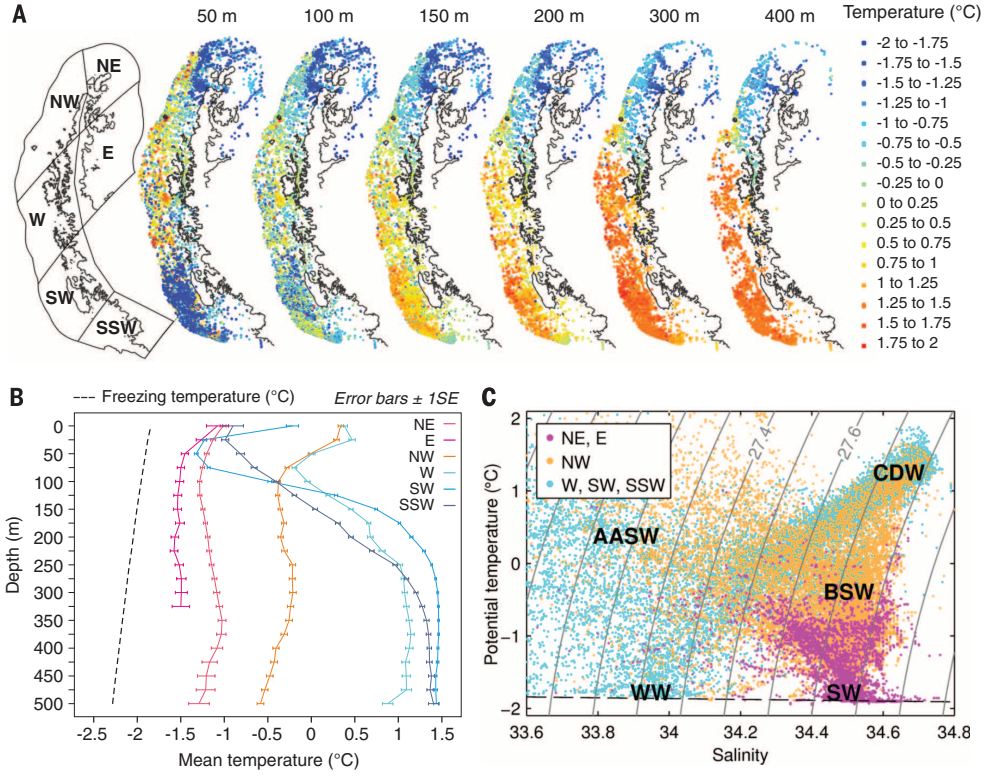



**Fig. 2. Ocean conditions surrounding the AP.** (**A**) In situ temperature of the ocean surrounding the AP at specific depths. The six regions are defined by east/west and by two-degree latitudinal bands, up to 100 km off the AP coast. (**B**) Mean in situ temperature profile in each region. The dashed line is the in situ freezing temperature. (**C**) Potential temperature–salinity diagram showing the different water masses in different regions, namely Shelf Water (SW), BSW, CDW, WW, and AASW. Gray lines are contours of surface density anomaly, and the dashed line is the freezing temperature.

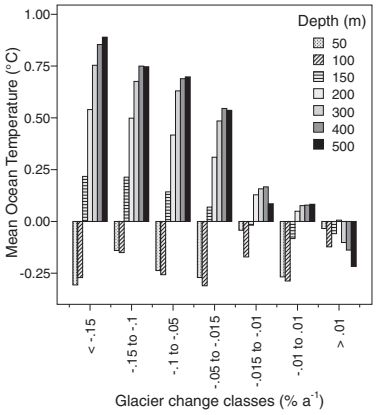



**Fig. 3. Mean overall glacier changes (binned) and mean in situ ocean temperature (within 5 km of glacier fronts) at specific depths.** Negative glacier change values signify retreat. The *x* axis reads from largest retreat rates on the left toward small changes and advances on the right.

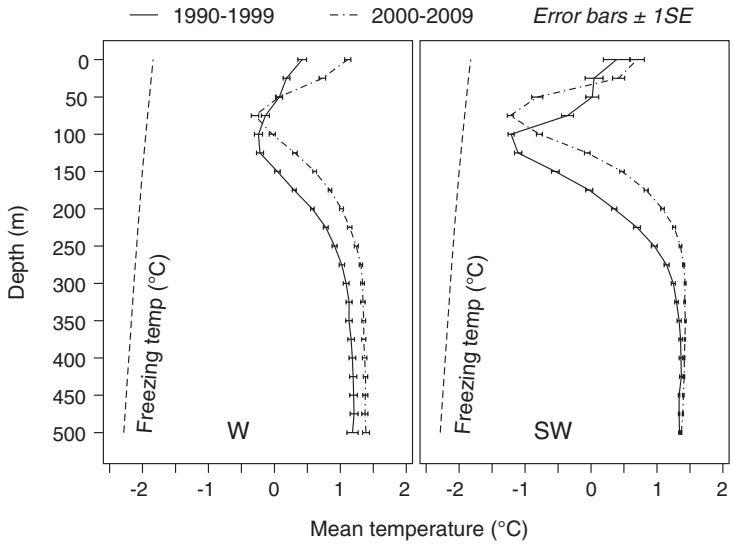



**Fig. 4. Mean in situ temperatures during the 1990s (solid line) and 2000s (dot-dash line) for repeated measurements in the west (W) and southwest (SW) regions.** The measurements were taken during cruises that occurred in the austral summer, with spatial coverage spanning the two regions.

and sea-ice formation over the Weddell Sea shelf (*28*). Ocean heat content in the Bellingshausen Sea is known to have increased, attributed to an increase in CDW upwelling onto the shelf, a decrease in heat loss to the atmosphere, and a slow warming of CDW offshore of the shelf (*29*, *30*). In common with changes on the Amundsen Sea shelf (*17*), the primary manifestation of the changing heat content is a change in the thickness of the deep CDW layer through a shoaling of the pycnocline. The BSW farther north, however, originates in a different climatic regime from CDW, where ocean temperatures are constrained near the surface freezing point by sea-ice processes, thus removing the potential impact of any temperature variability.

Although there are too few repeated ocean measurements before the 1990s to establish the significance of oceanic changes over the full period of glaciological data, there are sufficient observations of the northern Bellingshausen Sea to examine changes there since 1990. These observations reveal that the ocean was warmer on average in the 2000s than the 1990s, particularly at depths between 100 and 300 m in the southwest region (Fig. 4). During the late 1990s, a universal acceleration in glacier retreat occurred, apparent in all coastal regions except in the northwest (*8*) (fig. S2). The available oceanographic observations are therefore consistent with our hypothesis of ocean-driven glacier retreat.

We conclude that ocean temperatures below 100-m depth have been the predominant control on multidecadal glacier front behavior in the western AP. Glaciers abutting the warm, and warming, CDW regions in the Bellingshausen Sea have retreated, whereas those discharging into cooler BSW in the Bransfield Strait have not. The wide-scale regional ocean temperature pattern has existed since the earliest records, and the ocean heat content in regions dominated by CDW has increased since at least as long ago as the 1990s. Warming has primarily occurred at mid-depths (100 to 300 m) in these regions. Most important, ocean warming has occurred concurrently with a widespread acceleration in glacier retreat.

Ice shelves farther south in the Bellingshausen Sea are already recognized as being susceptible to ocean forcing (*13*, *14*, *31*), but this study shows that relatively warm coastal seas are also driving frontal retreat in 596 (90%) of the 674 marine-terminating glaciers farther north. Indeed, the climatic setting and marine-terminating nature of these AP glaciers mean that they share greater similarity to other "near-polar" environments where similar marine-terminating glaciers dominate (e.g., Greenland, Alaska, Patagonia, Svalbard, and some sub-Antarctic islands) than to the rest of Antarctica. Furthermore, our results emphasize the likely sensitivity of all such systems to changes in deep coastal waters and caution against assuming the dominance of atmospheric forcing, even where that warming is strong, as in the case of the AP. Our observations demonstrate clearly that simulations of glacier change over the past half-century that are driven solely by atmospheric climate [e.g., (*32*)] would fail to capture the most salient processes driving ice loss in the AP. It follows that predictive models employed to project future ice loss from glacial systems where marine-terminating glaciers abound will require coupling to oceanic, as well as atmospheric, forcing.

## REFERENCES AND NOTES


1. M. Huss, D. Farinotti, *The Cryosphere* **8**, 1261–1273 (2014).
2. G. J. Marshall, A. Orr, N. P. M. van Lipzig, J. C. King, *J. Clim.* **19**, 5388–5404 (2006).
3. M. Kunz *et al.*, *Geophys. Res. Lett.* **39**, 1–5 (2012).
4. N. E. Barrand *et al.*, *J. Geophys. Res. Earth Surf.* **118**, 315–330 (2013).
5. P. Holland *et al.*, *The Cryosphere* **9**, 1005–1024 (2015).
6. H. D. Pritchard, D. G. Vaughan, *J. Geophys. Res. Earth Surf.* **112**, F03S29 (2007).
7. E. M. Morris, D. G. Vaughan, in *Antarctic Peninsula Climate Variability: Historical and Paleoenvironmental Perspectives.*, Antarctic Research Series, vol. 79 (American Geophysical Union, 2003), pp. 61–68.
8. A. J. Cook, D. G. Vaughan, A. Luckman, T. Murray, *Antarct. Sci.* **26**, 614–624 (2014).
9. A. J. Cook, A. J. Fox, D. G. Vaughan, J. G. Ferrigno, *Science* **308**, 541–544 (2005).
10. J. C. Comiso, *J. Clim.* **13**, 1674–1696 (2000).
11. J. M. van Wessem *et al.*, *The Cryosphere* **10**, 271–285 (2016).
12. F. S. Paolo, H. A. Fricker, L. Padman, *Science* **348**, 327–331 (2015).
13. B. Wouters *et al.*, *Science* **348**, 899–903 (2015).
14. P. R. Holland, A. Jenkins, D. M. Holland, *J. Geophys. Res. Oceans* **115**, C05020 (2010).
15. H. D. Pritchard *et al.*, *Nature* **484**, 502–505 (2012).
16. A. Shepherd, D. J. Wingham, E. Rignot, *Geophys. Res. Lett.* **31**, L23402 (2004).
17. P. Dutrieux *et al.*, *Science* **343**, 174–178 (2014).
18. A. Luckman *et al.*, *Nat. Commun.* **6**, 8566 (2015).
19. T. P. Boyer *et al.*, *World Ocean Database 2013*, https://www.nodc.noaa.gov/OC5/WOD13/ (National Oceanic and Atmospheric Administration, 2013).
20. Materials and methods are available as supplementary materials on *Science* Online.
21. T. Whitworth III, W. Nowlin Jr., A. Orsi, R. Locarnini, S. Smith, *Deep Sea Res. Part I Oceanogr. Res. Pap.* **41**, 629–641 (1994).
22. E. E. Hofmann, J. M. Klinck, C. M. Lascara, D. A. Smith, in *Foundations for Ecological Research West of the Antarctic Peninsula*, vol. 70 (American Geophysical Union, 1996), pp. 61–81.
23. P. R. Holland, A. Jenkins, D. M. Holland, *J. Clim.* **21**, 2558–2572 (2008).
24. A. Jenkins, *J. Phys. Oceanogr.* **41**, 2279–2294 (2011).
25. P. Greisman, *Deep-Sea Res. A, Oceanogr. Res. Pap.* **26**, 1051–1065 (1979).
26. R. J. Motyka, L. Hunter, K. A. Echelmeyer, C. Connor, *Ann. Glaciol.* **36**, 57–65 (2003).
27. F. Straneo *et al.*, *Nat. Geosci.* **4**, 322–327 (2011).
28. T. S. Dotto, R. Kerr, M. M. Mata, C. A. Garcia, *J. Geophys. Res. Oceans* **121**, C01122 (2016).
29. D. G. Martinson, S. E. Stammerjohn, R. A. Iannuzzi, R. C. Smith, M. Vernet, *Deep Sea Res. Part II Top. Stud. Oceanogr.* **55**, 1964–1987 (2008).
30. S. Schmidtko, K. J. Heywood, A. F. Thompson, S. Aoki, *Science* **346**, 1227–1231 (2014).
31. L. Padman *et al.*, *J. Geophys. Res. Oceans* **117**, C01010 (2012).
32. J. Church *et al.*, *Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge Univ. Press, 2013).

ACKNOWLEDGMENTS

This work was largely supported through an AXA (global investment and insurance group) Research Fund Fellowship. We are grateful to the National Centers for Environmental Information for providing access to the World Ocean Database and to all those who have contributed ocean data from the region surrounding the Antarctic Peninsula. The glacier-change data reported in this paper are tabulated in a database in the supplementary materials. We are most grateful to the reviewers who gave recommendations for improving this paper.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/353/6296/283/suppl/DC1
Materials and Methods
Figs. S1 and S2
Table S1
Database S1
References (*33*)

4 December 2015; accepted 9 June 2016
10.1126/science.aae0017

EVOLUTIONARY COGNITION

# Ducklings imprint on the relational concept of "same or different"

**Antone Martinho III* and Alex Kacelnik***

**The ability to identify and retain logical relations between stimuli and apply them to novel stimuli is known as relational concept learning. This has been demonstrated in a few animal species after extensive reinforcement training, and it reveals the brain's ability to deal with abstract properties. Here we describe relational concept learning in newborn ducklings without reinforced training. Newly hatched domesticated mallards that were briefly exposed to a pair of objects that were either the same or different in shape or color later preferred to follow pairs of new objects exhibiting the imprinted relation. Thus, even in a seemingly rigid and very rapid form of learning such as filial imprinting, the brain operates with abstract conceptual reasoning, a faculty often assumed to be reserved to highly intelligent organisms.**

Relational concepts, such as "same" and "different," have been demonstrated in a few animal species, typically after extensive training (*1*, *2*). Relational concepts differ from other forms of categorical generalization. For instance, pigeons and bees can be trained to discriminate whether novel images contain humans or not (*3*), or whether novel paintings are by Monet or Picasso (*4*), by relying on the similarity between features of the training and of the novel stimuli. In relational concept learning, however, relative properties between training stimuli generate the relationship that has to be generalized to sets of novel stimuli (*5*).

The relations of "same" and "different" have been used to study relational concept learning in a few primates and birds (*6*), using a variety of protocols. For instance, in the identity matching to sample (IMTS) protocol, an animal sees a sample stimulus and subsequently chooses between two test stimuli, one of which is identical to the sample. Reinforcement can be contingent on responding to the identical one ("same") or to the alternative ("different"). Honey bees can learn this discrimination and even transfer a correct response to novel stimuli across sensory modalities (olfaction and visual texture) (*7*). The IMTS task requires learning the appropriate comparison between the working-memory representation of the sample and the currently perceived test stimuli, but it does not require interpreting an abstract relationship between perceived items and then reapplying the same relation to discriminate between sets of novel objects.

A different procedure, that isolates relational learning, involves presenting more than one stimulus as a sample, and then selecting, from between various sets of stimuli, the set that has

Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.
***Corresponding author. Email: antone.martinho@zoo.ox.ac.uk (A.M.); alex.kacelnik@zoo.ox.ac.uk (A.K.)**

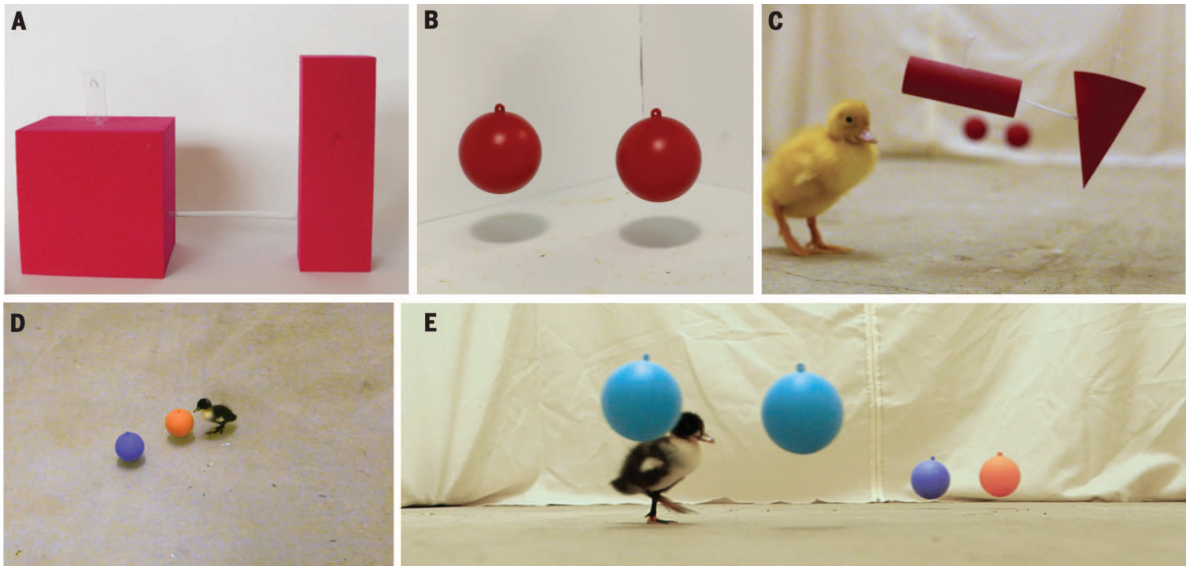



**Fig. 1. Imprinting and testing stimuli.** Newborn ducklings were first exposed to a pair of objects revolving about the center of a training arena, then tested with two novel pairs of objects. (**A**) Example of a "different shape" training stimulus pair. (**B**) Example of a "same color" training stimulus pair. After this exposure, the ducklings were tested for their preference between two novel stimulus pairs revolving in apposition. (**C**) A duckling trained with the set shown in (A) demonstrates its preference for a novel "different shape"' stimulus over an also-novel "same shape" stimulus. (**D**) A duckling trained with the stimulus pair shown in (B) approaches a novel "different color" stimulus pair—an incorrect response. (**E**) The same duckling later in the same trial correctly approaches and closely follows the novel "same color" stimulus.

the same internal relation as the sample set, a procedure called relational matching to sample (RMTS) (*1*, *8*). In RMTS, what has to be retained is a relation between stimuli, rather than a representation of a perceived stimulus. Primates (*9*), pigeons (*5*), parrots (*10*), and corvids (*1*) have succeeded in solving such problems, and it has been cogently argued that this may indicate their possession of the ability to reason by analogy (*8*).

Demonstrating the capacity for both identity and relational matching to sample has so far used reinforcement, and often extensive training, to allow subjects to infer the target relation. This contrasts with avian filial imprinting, a specialized form of unrewarded learning by which hatchlings acquire the ability to identify and then follow a parent or substitute parental object (*11*–*13*). As could be expected from its biological adaptive significance, imprinting is one of the fastest and most reliable forms of learning (*14*). In chicks and ducklings, high-fidelity imprinting responses can be acquired in a few minutes of unrewarded exposure to a stimulus (*15*).

Filial and sibling imprinting cannot be mediated just by snapshot representations of two-dimensional retinal images. Wood summarized the problem sharply: "Building an invariant object representation requires transforming patterns of retinal activity (view-specific information) into an abstract representation that is tolerant to retinal image changes and selective for a particular object (identity information)" (*16*). This is particularly relevant when considering that ducklings may benefit by recognizing not just their mother but also their group of siblings (*17*), because broods may have different degrees of heterogeneity. Evidence showing that young birds are sensitive to abstract qualities of stimuli, both spontaneously and through imprinting, supports this view and shows that imprinting is far richer, as a learning phenomenon, than had originally been envisaged (*18*–*21*). It is thus tempting to ask whether their abstraction abilities may extend to the more demanding phenomenon of relational concept representation, which has so far only been demonstrated through extensive reinforcement in species with advanced intelligence. To this effect, we modified the RMTS protocol to combine it with imprinting, as follows.

Following Bateson's (*22*) and Lickliter's (*23*) procedures for effective imprinting, we hatched domesticated mallard ducklings in the dark and kept them for 1 hour in a social group, with light, food, and water. They were then exposed for 25 min to a moving pair of sample objects, kept for a 30-min retention interval in the dark, and presented with two novel pairs of moving objects for 10 min (Fig. 1). Sample stimuli within the pair shown in the imprinting phase were either equal to each other in both color and shape, or different in one of these characteristics. Test stimuli in the preference test consisted of two stimuli pairs, composed of objects novel to the birds. In one test pair, the objects were equal in color and shape, and in the other they differed in either shape or color (Fig. 2). Detailed methods may be found in the supplementary materials.

In experiment 1, the imprinting phase stimulus pair consisted of two red solids, which were equal to each other in shape for group "same" and different for group "different" ($n$ = 36 ducklings for each group). The stimuli forming the two test pairs were also red, but one pair consisted of two novel, identical shapes and the other of two novel, different shapes. In experiment 2, the imprinting stimulus pair was two spheres, either of the same color or different colors ($n$ = 40 ducklings for each group), and test stimuli were also spherical but had novel colors, equal in one pair and different in the other.

To measure preference, the number of approaches undertaken by each duckling toward each test pair was scored twice, once by an

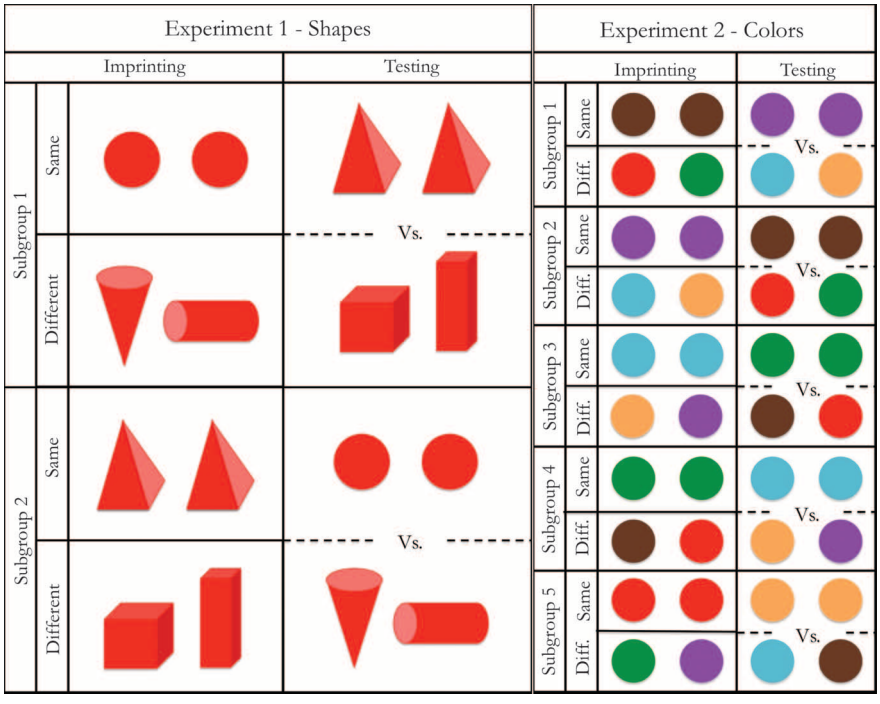



**Fig. 2. Experimental stimulus pairs.** Experiment 1 tested for responses between red objects that could differ in shape, and experiment 2 tested for responses between spherical objects that could differ in color, using the pairs illustrated here. In experiment 1, members of subgroup 1 were initially exposed to either two spheres or a pair formed by a cone and a cylinder and were then tested for preference between a pair of two pyramids and a pair formed by a prism and a cube, whereas members of subgroup 2 were imprinted on pairs of either two pyramids or a prism and a cube and were then tested on preference between a pair of spheres and a pair formed by a cone and a cylinder. In experiment 2, members of each subgroup were trained on either two identical spheres or a pair formed by two spheres of different colors, and all ducklings were tested for preference between two novel identical spheres and a pair formed by two novel differing spheres.

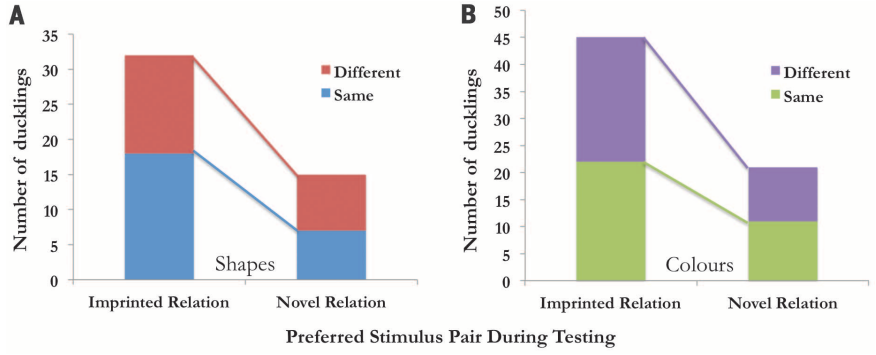



**Fig. 3. Ducklings' preferences for an imprinted relational concept.** The number of ducklings showing preference for the imprinted (left column) or alternative (right column) relation is shown for shape relations in experiment 1 (**A**) and the same for color relations in experiment 2 (**B**). Ducklings preferred the imprinted relation in both shape and color, regardless of whether the imprinted relation was "same" or "different."

independent scorer blind to each duckling's imprinting condition and to the study's hypothesis. Ducklings that were inactive during testing (fewer than five approaches) were excluded from analysis. Preferences were assessed via sign test, with sample size being the number of individual ducklings. Ducklings making more than half of their approaches toward a given stimulus were scored as having preferred it (see the methods in the supplementary materials). Video of sample trials (movie S1) is available in the supplementary materials.

Figure 3 shows the preference results. In experiment 1, out of a total of 47 active ducklings, 32 preferred the pair bearing their imprinted shape relation (two-tailed binomial test, $P = 0.02$). In experiment 2, out of 66 active ducklings, 45 preferred the stimulus pairs bearing their imprinted color relation (two-tailed binomial test, $P = 0.004$). Combining both results, out of 113 active ducklings, 77 preferred the relational concept, same or different, upon which they had imprinted (two-tailed binomial test, $P < 0.0001$).

The accuracy of our ducklings was comparable to, or better than, reinforced relational concept discrimination in primates (*24*) and crows (*1*). This finding supports a richer emerging view of the representation of information in the animal brain than is presently prevalent, in which even relatively simple learning systems do not process information just through the content of sensory signals but also by encoding higher-level, abstract aspects of stimulus analyses, already the target of neural network models designed to simulate such cognitive function (*25*). The ducklings' performance indicates that their brains may be prepared, not just to respond differentially to certain visual inputs, such as scrambled objects containing species-specific elements like legs or heads or virtual points that move in a biologically plausible coordination (*20*), but also to pick up abstract relational properties between elements of their sensory input and those elements' characteristics.

For young precocious birds, having this competence makes biological sense. For a duckling critically dependent on proximity to its mother and siblings, defining the attachment stimulus configuration as a library of sensory inputs and logical rules increases the likelihood that the mother and sibling group will be identified with high fidelity in spite of considerable variations in how they are perceived. The rules that may define the imprinted attachment target are likely to extend beyond properties of a single object such as color, shape, or symmetry, to include properties of object assemblies such as their informational entropy (*26*).

## REFERENCES AND NOTES


1. A. Smirnova, Z. Zorina, T. Obozova, E. Wasserman, *Curr. Biol.* **25**, 256–260 (2015).
2. E. A. Wasserman, *Psychol. Bull.* **113**, 211–228 (1993).
3. R. J. Herrnstein, D. H. Loveland, C. Cable, *J. Exp. Psychol. Anim. Behav. Process.* **2**, 285–302 (1976).
4. W. Wu, A. M. Moreno, J. M. Tangen, J. Reinhard, *J. Comp. Physiol. A Neuroethol. Sens. Neural Behav. Physiol.* **199**, 45–55 (2013).
5. T. R. Zentall, M. Galizio, T. S. Critchfied, *J. Exp. Anal. Behav.* **78**, 237–248 (2002).
6. A. A. Wright, J. S. Katz, *Behav. Processes* **72**, 234–254 (2006).
7. M. Giurfa, S. Zhang, A. Jenett, R. Menzel, M. V. Srinivasan, *Nature* **410**, 930–933 (2001).
8. R. G. Cook, E. A. Wasserman, *Psychon. Bull. Rev.* **14**, 1107–1114 (2007).
9. E. A. Wasserman, J. Fagot, M. E. Young, *J. Comp. Psychol.* **115**, 42–52 (2001).
10. I. M. Pepperberg, *Anim. Learn. Behav.* **15**, 423–432 (1987).
11. K. Lorenz, *J. Ornithol.* **83**, 289–413 (1935).
12. P. Bateson, *Anim. Behav.* **27**, 470–486 (1979).
13. P. Bateson, *Anim. Learn. Behav.* **7**, 259–262 (1979).
14. P. P. G. Bateson, *Biol. Rev. Camb. Philos. Soc.* **41**, 177–211 (1966).
15. P. Bateson, J. B. Jaeckel, *Anim. Behav.* **24**, 386–390 (1976).
16. J. N. Wood, *Dev. Sci.* **18**, 194–205 (2015).
17. E. H. Hess, D. B. Hess, *Psychon. Sci.* **14**, 129–130 (1969).
18. O. Rosa-Salva, L. Regolin, G. Vallortigara, *Dev. Sci.* **13**, 565–577 (2010).
19. G. Vallortigara, L. Regolin, F. Marconato, *PLOS Biol.* **3**, e208 (2005).
20. L. Regolin, L. Tommasi, G. Vallortigara, *Anim. Cogn.* **3**, 53–60 (2000).
21. R. Rugani, L. Regolin, G. Vallortigara, *Dev. Sci.* **13**, 790–797 (2010).
22. P. P. G. Bateson, A. A. P. Wainwright, *Behaviour* **42**, 279–290 (1972).
23. R. Lickliter, G. Gottlieb, *J. Comp. Psychol.* **101**, 40–46 (1987).
24. J. S. Katz, A. A. Wright, J. Bachevalier, *J. Exp. Psychol. Anim. Behav. Process.* **28**, 358–368 (2002).
25. P. Bateson, G. Horn, *Anim. Behav.* **48**, 695–715 (1994).
26. T. R. Zentall, E. A. Wasserman, O. F. Lazareva, R. K. Thompson, M. J. Rattermann, *Comp. Cogn. Behav. Rev.* **3**, 13–45 (2008).



## ACKNOWLEDGMENTS

We thank the Oxford University Farm for allowing us to use their ducklings for this work, and Poultry Manager M. Colley for providing duckling husbandry and technical help. P. Bateson, E. Wasserman, and H. Smithson gave us valuable advice and recommendations regarding our methods and interpretation. We are also grateful to E. Truswell, who assisted with some experimental trials, and M. Lewis, who acted as an independent "blind" scorer for data analysis. This research was funded with assistance from the University of Oxford Department of Zoology. Materials and methods and a video of sample trials may be found in the supplementary materials. Video recordings of training and testing can be made available on request. Data obtained from video coding are available as supplementary materials on *Science* Online (table S1). The experiment was devised jointly by A.M. and A.K. Experimentation, data collection, and analysis were performed by A.M. with the advice and support of A.K.. The manuscript was written jointly by A.M. and A.K., and figures were prepared by A.M. and edited by A.K.


## SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/353/6296/286/suppl/DC1
Materials and Methods
Table S1
Movie S1
Reference (*27*)

5 February 2016; accepted 17 May 2016
10.1126/science.aaf4247

BIODIVERSITY

# Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment

**Tim Newbold,[1,2]* Lawrence N. Hudson,[3] Andrew P. Arnell,[1] Sara Contu,[3] Adriana De Palma,[3,4] Simon Ferrier,[5] Samantha L. L. Hill,[1,3] Andrew J. Hoskins,[5] Igor Lysenko,[4] Helen R. P. Phillips,[3,4] Victoria J. Burton,[3] Charlotte W. T. Chng,[3] Susan Emerson,[3] Di Gao,[3] Gwilym Pask-Hale,[3] Jon Hutton,[1,6] Martin Jung,[7,8] Katia Sanchez-Ortiz,[3] Benno I. Simmons,[3,4] Sarah Whitmee,[2] Hanbin Zhang,[3] Jörn P. W. Scharlemann,[1,8] Andy Purvis[3,4]**

**Land use and related pressures have reduced local terrestrial biodiversity, but it is unclear how the magnitude of change relates to the recently proposed planetary boundary ("safe limit"). We estimate that land use and related pressures have already reduced local biodiversity intactness—the average proportion of natural biodiversity remaining in local ecosystems—beyond its recently proposed planetary boundary across 58.1% of the world's land surface, where 71.4% of the human population live. Biodiversity intactness within most biomes (especially grassland biomes), most biodiversity hotspots, and even some wilderness areas is inferred to be beyond the boundary. Such widespread transgression of safe limits suggests that biodiversity loss, if unchecked, will undermine efforts toward long-term sustainable development.**

Land use and related pressures have been the main drivers of terrestrial biodiversity change (*1*) and are increasing (*2*). Biodiversity has already experienced widespread large net losses (*3*), potentially compromising its contribution to resilient provision of ecosystem functions and services, such as biomass production and pollination, that underpin human well-being (*4*–*7*). Species-removal experiments suggest that loss of ecosystem function accelerates with ongoing species loss (*5*), implying that there may be thresholds beyond which human intervention is needed to ensure adequate local ecosystem function (*8*, *9*). The loss of 20% of species—which affects ecosystem productivity as strongly as other direct drivers (*5*)—is one possible threshold, but it is unclear by which

mechanism species richness affects ecosystem function and whether there are direct effects or only effects on resilience of function (*6*, *7*). Whereas this proposed safe limit comes from studies of local ecosystem health, the Planetary Boundaries framework (*8*, *9*) considers longer-term maintenance of function over much larger (biome to global) scales. At these temporal and spatial scales, the maintenance of function depends on functional diversity—the ranges and abundances of the functional traits of the species present (*8*, *10*). Because direct functional trait data are lacking, the Biodiversity Intactness Index [BII; the average abundance of originally present species across a broad range of species, relative to abundance in an undisturbed habitat (*11*)] is suggested as the best metric (*8*, *9*). The safe limit is placed at a precautionary 10% reduction in BII, but it might be as high as a 70% reduction (*9*).

A key uncertainty when estimating safe limits concerns the value of species not present in the undisturbed ecosystem. Such species could benefit ecosystem functioning, have no effect (as assumed by the BII), or even impair it (*12*–*15*). Most models estimating net human impacts on biodiversity (*3*, *16*) treat novel and originally present species as functionally equivalent, whereas experimental studies manipulate species originally present (*17*).

Given the possibly severe consequences of transgressing safe biodiversity limits, global assessments of relevant metrics are needed urgently. Data limitations have hampered efforts to date; BII has so far only been estimated, from expert opinion, for seven southern African countries (*11*). More recently, we combined global models linking land-use pressures to local biodiversity with global land-use maps. We estimated that net reductions in local species richness exceeded 20% across 28% of the world's land surface by 2005, whereas 48.7% of land had seen net reductions in total abundance of ≥10% (*3*). However, our projections of net effects did not account for any reductions of originally present diversity that were offset by an influx of novel species (*18*), as well as being at too coarse a scale (~50 $km^2$) to be relevant for local ecosystem functioning and decision-making. Furthermore, we did not analyze the spatial distribution of the transgression of proposed safe limits.

[1]United Nations Environment Programme World Conservation Monitoring Centre, 219 Huntingdon Road, Cambridge CB3 0DL, UK. [2]Centre for Biodiversity and Environment Research, Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, UK. [3]Department of Life Sciences, Natural History Museum, Cromwell Road, London SW7 5BD, UK. [4]Department of Life Sciences, Imperial College London, Silwood Park, London SL5 7PY, UK. [5]Commonwealth Scientific and Industrial Research Organization (CSIRO) Land and Water Flagship, Canberra, Australian Capital Territory 2601, Australia. [6]Luc Hoffmann Institute, World Wildlife Fund (WWF) International, 1196 Gland, Switzerland. [7]Center for Macroecology, Evolution and Climate, Natural History Museum of Denmark, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen Ø, Denmark. [8]School of Life Sciences, University of Sussex, Brighton BN1 9QG, UK.
***Corresponding author. Email: t.newbold@ucl.ac.uk**

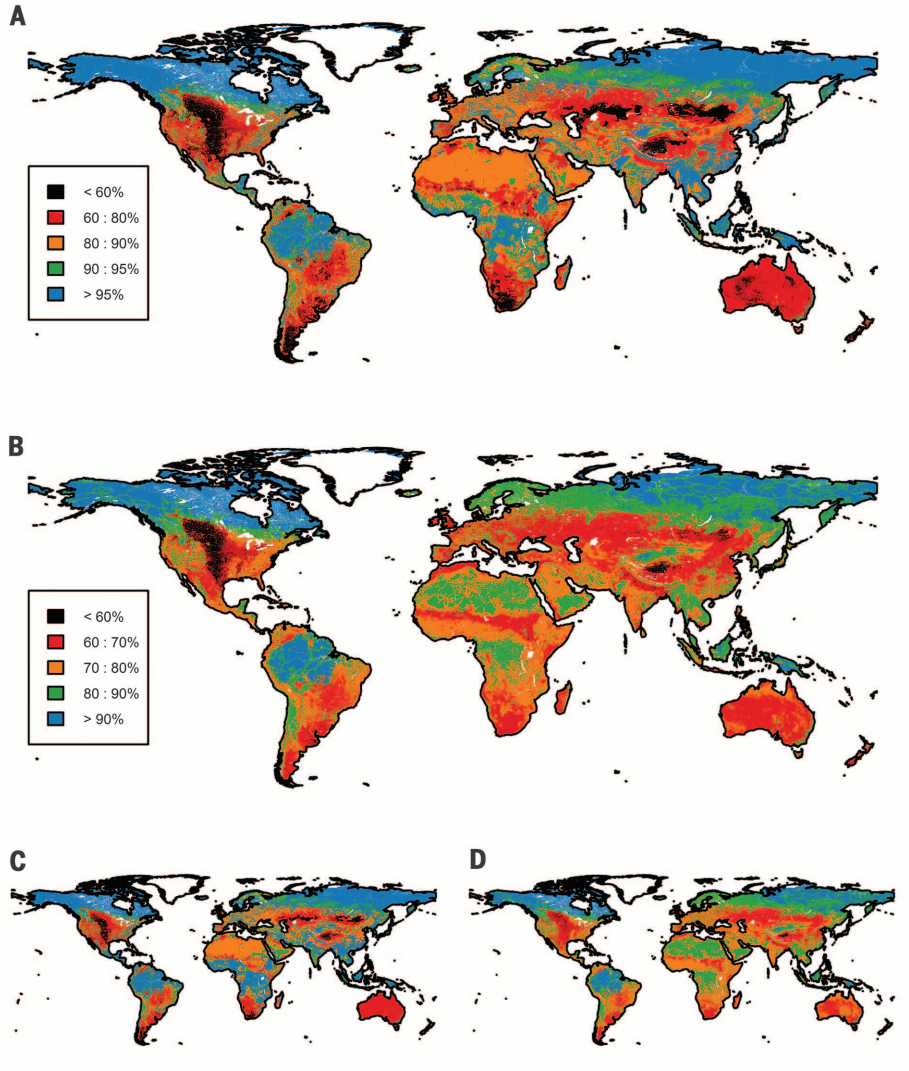



**Fig. 1. Biodiversity intactness of ecological assemblages.** (**A**) Total abundance of species occurring in primary vegetation (BII). (**B**) Richness of species occurring in primary vegetation. (**C**) and (**D**) correspond to (A) and (B), respectively, and have the same legend values but include species not present in primary vegetation.

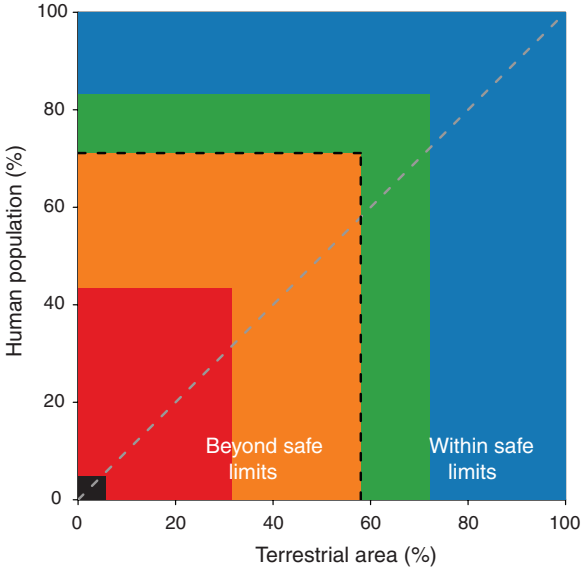



**Fig. 2. Terrestrial area and human population at different levels of BII.** Biodiversity intactness increases from bottom left to top right and has the same color scheme as that of Fig. 1. The dashed black line shows the position of the planetary boundary (*9*): Only areas to the right and human population above this line (shaded green and blue) are within the proposed safe operating space. If human population were distributed randomly with respect to BII, the corners of the boxes would align with the dashed gray line; the extent to which the corners lie above this line indicates the strength of the bias in human populations toward less intact areas.

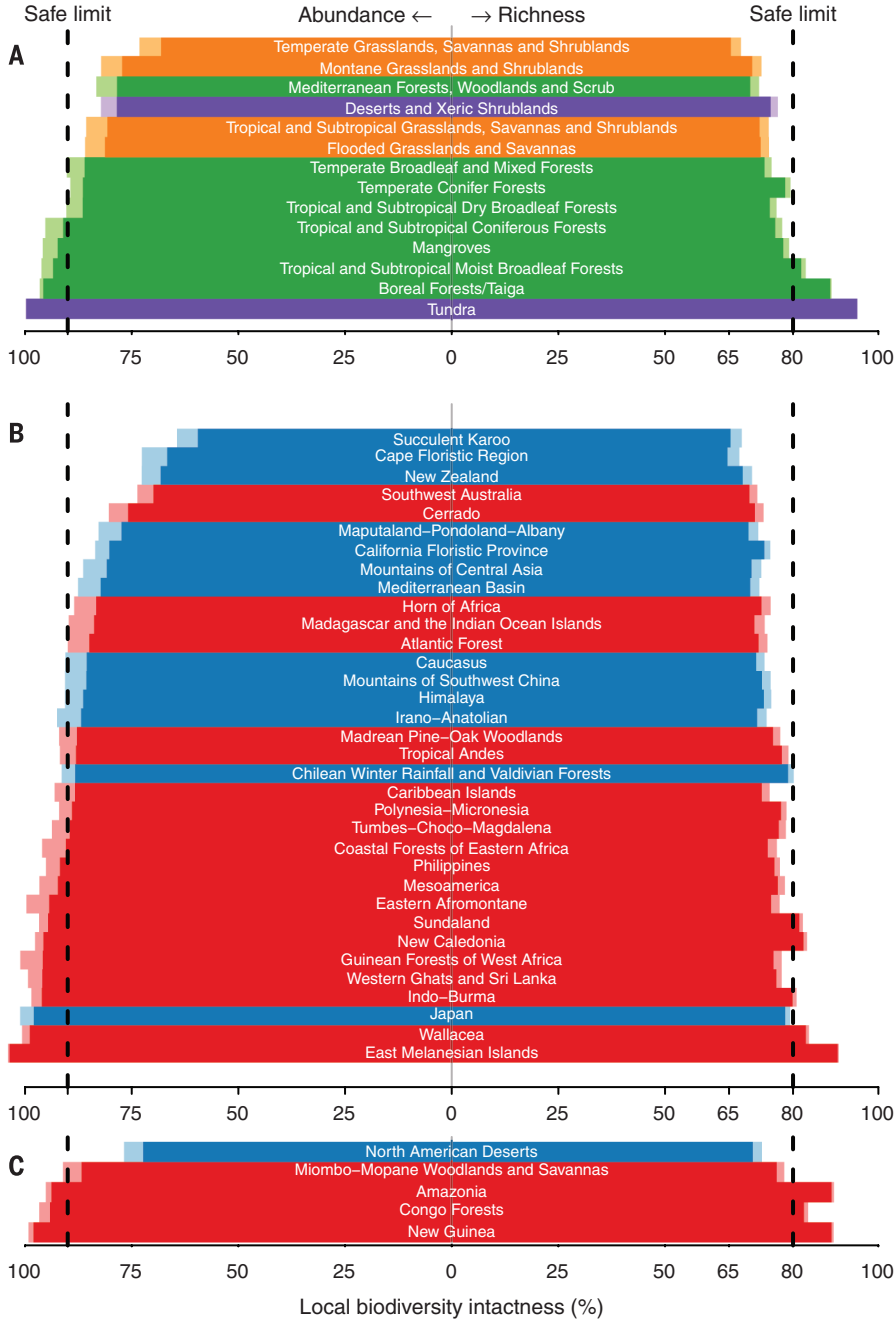



**Fig. 3. Biodiversity intactness for biomes, biodiversity hotspots, and high biodiversity wilderness areas.** (**A** to **C**) Biodiversity intactness in terms of total abundance (BII; solid bars on left) and species richness (solid bars on right) in each of (A) 14 terrestrial biomes, (B) 34 biodiversity hotspots, and (C) five high biodiversity wilderness areas. Translucent bars indicate the corresponding relative biodiversity values if novel species are treated as equivalent to those originally present (these numbers can surpass 100% because gains may outnumber losses). Bars in (A) are colored by major biome type (orange, grasslands; green, forests; purple, other), whereas bars in (B) and (C) are colored according to whether they are in the temperate (blue) or tropical (red) realms.

Here, we present fine-scale (~1 $km^2$) global estimates of how land-use pressures have affected the numbers of species and individuals found in samples from local terrestrial ecological assemblages (*19*). To explore different assumptions about novel species, we estimated both overall net change (correct if novel species contribute fully to ecosystems) and—using estimates of species turnover among land uses to exclude novel species—change in species originally present (correct if novel species play no role). We asked how much of the Earth's land surface is already "biotically compromised" (exceeds the boundaries of 10% loss of abundance or 20% loss of species). We focused on results for the relative abundance of originally present species (BII) because this is the measure suggested in the Planetary Boundaries framework (*9*). We estimated average losses per biome because of the suggested importance of biomes for the functioning of the whole Earth system (*8*, *9*), and to assess possible consequences for people—assuming that many biodiversity-regulated ecosystem services operate locally—we quantified the geographical congruence between biodiversity reduction and human population. We also assessed the biotic integrity of areas identified as particularly important for conservation (although the proposed planetary boundary in terms of BII may not always be relevant for areas much smaller than biomes and probably needs to vary depending on the sensitivity of the biota). First, Conservation International's "biodiversity hotspots"—areas rich in endemic species but with high levels of habitat loss—have been suggested as urgent conservation priorities (*20*). Because these areas were identified reactively (*20*) with a criterion of 70% loss of primary vegetation, we expect them to have lower biodiversity intactness than average. For comparison, we also estimated the biodiversity intactness of Conservation International's high biodiversity wilderness areas, which also meet the criterion of high species endemism but retain 70% of their natural habitat, and so present more opportunity for proactive conservation (*20*).

We modeled how sampled richness and abundance respond to land-use pressures using data from the PREDICTS (Projecting Responses of Ecological Diversity in Changing Terrestrial Systems) database (*21*). These data consisted of 2,382,624 records (fig. S1) [nearly twice as many as our earlier, coarser-scale analyses (*3*)] of the abundance (1,888,784 records) or else presence/absence of 39,123 species at 18,659 sites. The hierarchical mixed-effects models we used considered four pressure variables—land use, land-use intensity, human population density, and proximity to the nearest road—as fixed effects (figs. S2 and S3), whereas random effects accounted for among-study differences in sampling (methods, effort, and focal taxonomic groups) and for the spatial arrangement of sampled sites within studies (supplementary materials, materials and methods). We had insufficient data to fit separate models for each biome or clade. Responses may vary taxonomically or geographically, although our earlier analyses showed no significant differences among plants, invertebrates, and vertebrates and suggested limited variation among biomes (*3*). As more data become available, future analyses will be better able to reflect any differences in response. We combined the models of species richness and total abundance with models of species turnover among land uses [based on (*22*), but adapted to reflect asymmetric differences among land uses] to discount the fraction of species absent in nonprimary habitat (supplementary materials, materials and methods). To map modeled responses, we used global pressure data for the year 2005 at a resolution of 30 arc sec (~1 $km^2$). We used land-use estimates for 2005 (*23*) and estimated land-use intensity as in (*3*); human

population (for the year 2000) came from (*24*), and proximity to nearest road came from (*25*). Values of the response variables are always expressed relative to an intact assemblage undisturbed by humans and therefore do not rely on estimates of absolute abundance or species richness, which vary widely among biomes and taxa.

Our map of terrestrial BII (Fig. 1A and fig. S4) suggests that the average local abundance of originally present species (*11*) globally has fallen to 84.6% [95% confidence interval (CI), 82.2 to 91.6%] of its value in the absence of human land-use effects, which is probably below the value (90%) proposed as a safe limit (*9*). Considering net changes in abundance, as in (*3*), assuming that novel species contribute fully to ecosystem function, global average abundance has fallen to 88.0% (95% CI, 83.5 to 94.8%) of its value before human effects.

Assuming that only originally present species contribute to ecosystem function, most of the world's land surface is biotically compromised in terms of BII (58.1% of terrestrial area; 95% CI, 40.4 to 70.2%) (Fig. 1A) and within-sample richness of originally present species (62.4%; 95% CI, 20.0 to 72.7%) (Fig. 1B). If the proposed boundaries are broadly correct, ongoing human intervention may be needed to ensure delivery of ecosystem functions across most of the world (*5*). The proposed planetary boundary for BII (*9*) had uncertainty ranging from 30 to 90%; the proportion of the land surface exceeding the boundary varies widely across this range (fig. S5), highlighting the urgent need for better understanding of how BII relates to Earth-system functioning (*9*). Assuming that novel species contribute as much to ecosystems as originally present species, we estimate the safe limit for total abundance to have been crossed in 48.4% (95% CI, 30.9 to 66.5%) of land (Fig. 1C) and that for within-sample species richness in 58.4% (95% CI, 21.8 to 75.0%) (Fig. 1D). If novel species impair ecosystem function (rather than benefit it or have no effects), then all of these estimates will be too optimistic. Most people (71.4%) live in biotically compromised areas, as judged with BII (Fig. 2), although uncertainty in this result was high (95% CI, 8.7 to 92.4%). There is growing evidence that access to high-biodiversity areas benefits people's physical and psychological well-being (*26*, *27*), although uncertainty remains over which aspects of biodiversity are important.

The biodiversity impact of land-use pressures varies among biomes (Fig. 3A and table S2): grasslands are most affected, and tundra and boreal forests are least affected. Our BII estimates suggest 9 of the 14 terrestrial biomes (95% CI, 4 to 12) have on average transgressed safe limits for biodiversity (Fig. 3A), although this number drops to seven (95% CI, 1 to 12) if novel species are included. The BII limit has been crossed in 22 of 34 terrestrial biodiversity hotspots (95% CI, 7 to 31) (Fig. 3B and table S3) (*28*); this figure falls to 12 (95% CI, 5 to 32) if novel species are included, again highlighting the need to understand their effects on ecosystem function. Given that biodiversity hotspots were identified partly on the basis of widespread historical habitat loss (*20*), their low average BII is unsurprising, although our results suggest that at least some hotspots might stay within safe ecological limits if future land conversion is reduced. In contrast, three out of the five high biodiversity wilderness areas, which were identified for conservation proactively because the habitat is still relatively intact (*20*), have not experienced average losses of local biodiversity (BII) that cross the planetary boundaries (95% CI, 2 to 4) (four out of five if novel species are included; 95% CI, 2 to 5) (Fig. 3C and table S4). Results concerning which areas have crossed proposed planetary boundaries were generally consistent between the richness- and abundance-based biodiversity measures (Fig. 3 and tables S2 to S4).

Our models suggest a generally smaller impact of land use on BII than that in a previous study (*11*). This might reflect differences in taxonomic coverage, but there are also two reasons why our results may overestimate BII. First, we ignore lagged responses. Second, our models use sites in primary vegetation as a baseline because historical data are so rare (*3*, *11*); these sites will often have experienced some human impact. Nevertheless, because our models are global, their baseline is not biome- or region-specific and they do not rely on data from minimally affected land use from heavily modified landscapes, where such conditions do not exist. Our data have good coverage of taxa and biomes (fig. S1), but the density of sampling is inevitably uneven. Biomes that are particularly underrepresented, relative to their global ecosystem productivity, are boreal forests, tundra, flooded grasslands, and savannas and mangroves (fig. S1), meaning that less confidence can be placed in the results for these biomes. The data probably also underrepresent soil and canopy species. The estimate of land area biotically compromised in terms of species richness is much higher than our previous assessment (58.4 versus 28.4%, although the CIs overlap), but the estimates based on total abundance are almost identical (48.4 versus 48.7%) (*3*). The discrepancy for species richness is because of a stronger modeled interaction here between land use and human population density (fig. S3) and because we include the effect of roads and the interaction between roads and land use, which were omitted from the projections in (*3*).

The Sustainable Development Goals adopted in September 2015 (*29*) aim to improve human well-being while protecting, restoring, and sustainably using terrestrial ecosystems. Our results highlight the magnitude of the challenge. Exploitation of terrestrial systems has been vital for human development throughout history (*30*), but the cost to biosphere integrity has been high. Slowing or reversing the global loss of local biodiversity will require preserving the remaining areas of natural (primary) vegetation and, so far as possible, restoring human-used lands to natural (secondary) vegetation. Such an outcome would be beneficial for biodiversity, ecosystems, and—at least in the long term—human well-being.

## REFERENCES AND NOTES


1. J.-C. Vié, C. Hilton-Taylor, S. N. Stuart, *Wildlife in a Changing World: An Analysis of the 2008 IUCN Red List of Threatened Species* (IUCN, 2009).
2. D. P. Tittensor *et al.*, *Science* **346**, 241–244 (2014).
3. T. Newbold *et al.*, *Nature* **520**, 45–50 (2015).
4. B. J. Cardinale *et al.*, *Nature* **486**, 59–67 (2012).
5. D. U. Hooper *et al.*, *Nature* **486**, 105–108 (2012).
6. Y. Hautier *et al.*, *Science* **348**, 336–340 (2015).
7. T. H. Oliver *et al.*, *Trends Ecol. Evol.* **30**, 673–684 (2015).
8. G. M. Mace *et al.*, *Glob. Environ. Change* **28**, 289–297 (2014).
9. W. Steffen *et al.*, *Science* **347**, 1259855 (2015).
10. S. Díaz, M. Cabido, *Trends Ecol. Evol.* **16**, 646–655 (2001).
11. R. J. Scholes, R. Biggs, *Nature* **434**, 45–49 (2005).
12. P. M. Vitousek, *Oikos* **57**, 7–13 (1990).
13. J. M. Levine *et al.*, *Proc. Biol. Sci.* **270**, 775–781 (2003).
14. L. Pejchar, H. A. Mooney, *Trends Ecol. Evol.* **24**, 497–504 (2009).
15. M. Vilà *et al.*, *Ecol. Lett.* **14**, 702–708 (2011).
16. L. Gibson *et al.*, *Nature* **478**, 378–381 (2011).
17. M. Vellend *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 19456–19459 (2013).
18. C. D. Thomas, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 19187–19188 (2013).
19. Materials and methods are available as supplementary materials on *Science* Online.
20. T. M. Brooks *et al.*, *Science* **313**, 58–61 (2006).
21. L. N. Hudson *et al.*, *Ecol. Evol.* **4**, 4701–4735 (2014).
22. T. Newbold *et al.*, *Ecography* 10.1111/ecog.01932 (2016).
23. A. J. Hoskins *et al.*, *Ecol. Evol.* **6**, 3040–3055 (2016).
24. CIESIN, IFPRI, CIAT, *Global Rural-Urban Mapping Project, Version 1 (GRUMPv1): Population Density Grid* (NASA SEDAC, Palisades, 2011); http://sedac.ciesin.columbia.edu/data/dataset/grump-v1-population-density.
25. CIESIN, ITOS, University of Georgia, *Global Roads Open Access Data Set, Version 1 (gROADSv1)* (NASA SEDAC, 2013); http://dx.doi.org/10.7927/H4VD6WDCT.
26. R. A. Fuller, K. N. Irvine, P. Devine-Wright, P. H. Warren, K. J. Gaston, *Biol. Lett.* **3**, 390–394 (2007).
27. P. A. Sandifer, A. E. Sutton-Grier, B. P. Ward, *Ecosyst. Serv.* **12**, 1–15 (2015).
28. N. Myers, R. A. Mittermeier, C. G. Mittermeier, G. A. B. da Fonseca, J. Kent, *Nature* **403**, 853–858 (2000).
29. United Nations, *UN General Assembly Resolution A/RES/70/1* (2015); www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E.
30. E. C. Ellis *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 7978–7985 (2013).


## ACKNOWLEDGMENTS

We thank the hundreds of researchers who contributed their data to the PREDICTS project and G. Mace and N. Burgess for helpful comments on an earlier draft. This work was funded by the UK Natural Environment Research Council (NERC, grants NE/J011193/2 and NE/L002515/1), the Biotechnology and Biological Sciences Research Council (grant BB/F017324/1), a Hans Rausing Ph.D. scholarship, and the United Nations Environment Programme World Conservation Monitoring Centre. This is a contribution from the Imperial College Grand Challenges in Ecosystem and the Environment Initiative and the Sussex Sustainability Research Programme. PREDICTS is endorsed by the Group on Earth Observations Biodiversity Observation Network (GEO-BON). The underlying biodiversity data can be downloaded from the Natural History Museum's Data Portal (http://dx.doi.org/10.5519/0073893).

## SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/353/6296/288/suppl/DC1
Materials and Methods
Figs. S1 to S7
Tables S1 to S7
References (*31*–*457*)

11 January 2016; accepted 6 June 2016
10.1126/science.aaf2201

NEURODEVELOPMENT

# Return to quiescence of mouse neural stem cells by degradation of a proactivation protein

Noelia Urbán,[1]* Debbie L. C. van den Berg,[1] Antoine Forget,[2,3] Jimena Andersen,[1]† Jeroen A. A. Demmers,[4] Charles Hunt,[1] Olivier Ayrault,[2,3] François Guillemot[1]*

**Quiescence is essential for long-term maintenance of adult stem cells. Niche signals regulate the transit of stem cells from dormant to activated states. Here, we show that the E3-ubiquitin ligase Huwe1 (HECT, UBA, and WWE domain–containing 1) is required for proliferating stem cells of the adult mouse hippocampus to return to quiescence. Huwe1 destabilizes proactivation protein Ascl1 (achaete-scute family bHLH transcription factor 1) in proliferating hippocampal stem cells, which prevents accumulation of cyclin Ds and promotes the return to a resting state. When stem cells fail to return to quiescence, the proliferative stem cell pool becomes depleted. Thus, long-term maintenance of hippocampal neurogenesis depends on the return of stem cells to a transient quiescent state through the rapid degradation of a key proactivation factor.**

Stem cells contribute to tissue homeostasis by generating new differentiated cells. Adult stem cells can enter a reversible state of quiescence that protects the cells from damage and the population from depletion. Niche signals determine the balance between quiescent and activated states. Excessive quiescence leads to too few differentiated progeny, whereas excessive proliferation exhausts the stem cell population (*1*).

Neural stem cells (NSCs) in the dentate gyrus (DG) of the mouse hippocampus generate new granule neurons that integrate into the hippocampal circuit to modulate mood and memory (*2*, *3*). Niche signals control expression of the transcription factor Ascl1 (achaete-scute family bHLH transcription factor 1), which in turn directs NSC proliferation (*4*). To identify factors that regulate Ascl1, we characterized proteins that coimmunoprecipitate with Ascl1 in cultured mouse NSCs using mass spectrometry. We found that Huwe1 (HECT, UBA, and WWE domain–containing 1), a HECT domain E3 ubiquitin ligase associated with idiopathic intellectual disability and schizophrenia (*5*, *6*), interacts with Ascl1 (fig. S1). We generated embryonic telencephalon- and adult hippocampus–derived NSCs in which *Huwe1* is expressed and can be inactivated by Cre recombinase (fig. S2) (*7*). Inactivation of *Huwe1* resulted in an accumulation of Ascl1 protein and an extension of its half-life from 38 to 121 min (Fig. 1, A to C, and fig. S2, B, E, and G), whereas proteins destabilized by Huwe1 in other tissues were not affected (fig. S2C) (*8*–*10*). Ascl1 is degraded by the proteasome in NSCs (fig. S2D), and silencing of *Huwe1* with short hairpin RNAs (shRNAs) decreased the extent of polyubiquitinylation of Ascl1 (Fig. 1D). Therefore, Huwe1 promotes the proteasomal degradation of Ascl1 protein.

Huwe1 is expressed throughout the brain, including in hippocampal NSCs and their progeny in the subgranular zone of the DG (fig. S3). To study *Huwe1* function in these cells, we generated mice in which administration of the small molecule tamoxifen inactivates the *Huwe1* gene and initiates yellow fluorescent protein (YFP) expression in hippocampal NSCs [*Huwe1*$^{fl}$;*GLAST-CreERT2*; *Rosa-Stop-YFP* mice (*11*), called *Huwe1cKO* mice hereafter]. One month after tamoxifen administration, the intensity of Ascl1 immunolabel was enhanced in cells of the subgranular zone of *Huwe1cKO* mice compared with controls (Fig. 1, E, F, and G; and fig. S4). The number of GFAP$^+$ (glial fibrillary acidic protein–positive) radial NSCs expressing Ascl1 was also increased (Fig. 1H). We observed no difference in the expression of other known Huwe1 substrates (fig. S5). Thus, Huwe1 regulates Ascl1 stability in hippocampal NSCs. Because Ascl1 promotes NSC activation in the hippocampus (*4*), up-regulation of Ascl1 in *Huwe1cKO* mice might stimulate NSC proliferation. Indeed, a higher proportion of NSCs in the DG of *Huwe1cKO* mice were cycling at postnatal day 90 (P90) (Fig. 2, A and B). Thus, *Huwe1* suppresses hippocampal NSC proliferation in wild-type mice.

*Huwe1cKO* mice also had too few intermediate progenitors and neuroblasts, and the remaining cells ectopically expressed Ascl1 (Fig. 2C and fig. S6). The deletion of *Huwe1* did not induce a switch toward gliogenesis, and intermediate progenitors were most likely eliminated through apoptosis (figs. S7 and S8). We suggest that persistence of Ascl1 protein in progenitors lacking *Huwe1* maintains the proliferative state of NSCs and prevents differentiation of early intermediate progenitors.

To study the role of the interaction between Huwe1 and Ascl1 in the regulation of quiescence, we labeled quiescent NSCs by means of prolonged exposure to BrdU followed by a chase (label-retention assay) (*12*). We then inactivated *Huwe1* and analyzed the mice 3 weeks later (Fig. 3A). The numbers of BrdU-retaining progenitors were not significantly different in *Huwe1cKO* and control mice, indicating that the loss of *Huwe1* did not lead to premature activation of quiescent stem cells, which would result in BrdU dilution (Fig. 3B and fig. S9, A to F). Thus, *Huwe1* is not required to maintain NSCs in quiescence.

To determine whether *Huwe1* is required for proliferating NSCs to return to quiescence, we marked cells exiting the cell cycle in the absence of *Huwe1* by first inactivating *Huwe1* and then performing a BrdU label–retention assay (Fig. 3C). BrdU-retaining radial cells in the subgranular zone of control mice were quiescent NSCs and not astrocytes because they did not express the astrocytic marker S100β (fig. S9H). There were fewer BrdU-retaining NSCs in *Huwe1cKO* mice than in control mice (Fig. 3D and fig. S9, I and J), indicating that without *Huwe1*, fewer NSCs returned to quiescence. We could not directly examine the divisions of *Huwe1cKO* NSCs by means of in vivo clonal analysis (*13*) because the low dose of tamoxifen required for this analysis was not sufficient to delete the *Huwe1*$^{fl}$ mutant allele (fig. S10). To directly assess whether *Huwe1* is required in proliferating NSCs for their return to quiescence, we marked instead a cohort of proliferating cells with a pulse of EdU and identified the fractions of NSCs that had either exited or reentered the cell cycle 24 hours later by double labeling for EdU and Ki67 (Fig. 3E and fig. S11). In control mice, 23.4% of EdU$^+$ NSCs were negative for Ki67, suggesting that they had returned to quiescence after cycling and incorporating EdU (Fig. 3, F and G). In *Huwe1cKO* mice, only 2.6% of EdU$^+$ NSCs were negative for Ki67, indicating that almost all *Huwe1* mutant NSCs had reentered the cell cycle (Fig. 3, F and G). Thus, elimination of the proactivation factor Ascl1 from proliferating NSCs by Huwe1 in wild-type mice drives the cells into quiescence.

The long-term consequence of excessive proliferation of hippocampal NSCs in *Huwe1cKO* mice was examined 5 months after *Huwe1* deletion, at P210 (Fig. 3H). The overall number of NSCs was unchanged, confirming that *Huwe1* is not required for the maintenance of the predominant quiescent NSC population (Fig. 3, I and J). In contrast, the number of proliferating NSCs was reduced (2.4 ± 0.1% Ki67$^+$ NSCs in control mice; 0.3 ± 0.3% in *Huwe1cKO* mice) (Fig. 3K), indicating that *Huwe1* is required for the long-term maintenance of the proliferative NSC population in the hippocampus. This result also shows that stem cells that have proliferated and returned

[1]The Francis Crick Institute–Mill Hill Laboratory, NW7 1AA, London, UK. [2]Institut Curie, PSL Research University, CNRS UMR 3347, INSERM U1021, 91405, Orsay, France. [3]Université Paris Sud, Université Paris Saclay, CNRS UMR 3347, INSERM U1021, 91405, Orsay, France. [4]Proteomics Center, Erasmus MC, Wytemaweg 80, 3015 CN Rotterdam, Netherlands.
*Corresponding author. Email: noelia.urban@crick.ac.uk (N.U.); francois.guillemot@crick.ac.uk (F.G.). †Present address: Department of Psychiatry and Behavioral Sciences, Center for Sleep Sciences and Medicine, Stanford University School of Medicine, Stanford, CA, USA.

to quiescence are required to replenish the proliferative stem cell pool (fig. S12).

Ascl1 activates the transcription of several cell-cycle regulators in NSCs (*4*, *14*). *Huwe1*-deficient NSCs showed higher expression of *CcnD1* (*Cyclin D1*) and *CcnD2* (*Cyclin D2*), two targets of Ascl1 (figs. S13A and S14). The elevation of *CcnD1* and *CcnD2* expression in *Huwe1*-mutant NSCs was due to the accumulation of Ascl1 because it was abolished after *Ascl1* knockdown or deletion (figs. S13D and S14A). The increase in CcnD1 expression in *Huwe1cKO* mice was seen in quiescent NSCs and to a greater extent in proliferating NSCs (Fig. 4, B to E, and fig. S14F). Thus, stabilization of Ascl1 in NSCs lacking *Huwe1* promotes cell cycle reentry by inducing the expression of *CcnD* genes.

Posttranscriptional regulation controls stem cell activity, alongside transcriptional and epigenetic mechanisms (*15*). In the embryonic nervous system, Huwe1 promotes cell cycle exit and neuronal differentiation of progenitors by destabilizing N-myc (*7*). In the adult brain, we show here that Huwe1 targets the proactivation factor Ascl1 to promote the return of proliferating hippocampal NSCs to a resting state. Regulation of Ascl1 alone is not sufficient to promote quiescence exit, suggesting that additional signals are required to stimulate stem cell activity. Most NSCs continue to divide once activated and are eventually lost, thus contributing to the rapid attrition of the stem cell population over time (*16*). However, *Huwe1* promotes the return to a resting state of a minority of dividing NSCs, which is essential for the long-term maintenance of the diminishing pool of proliferating stem cells (fig. S12). Our results suggest that proliferating stem cells that return to quiescence form a pool of temporarily resting cells that is distinct from the main dormant pool and is the main contributor to neurogenesis in the adult hippocampus.

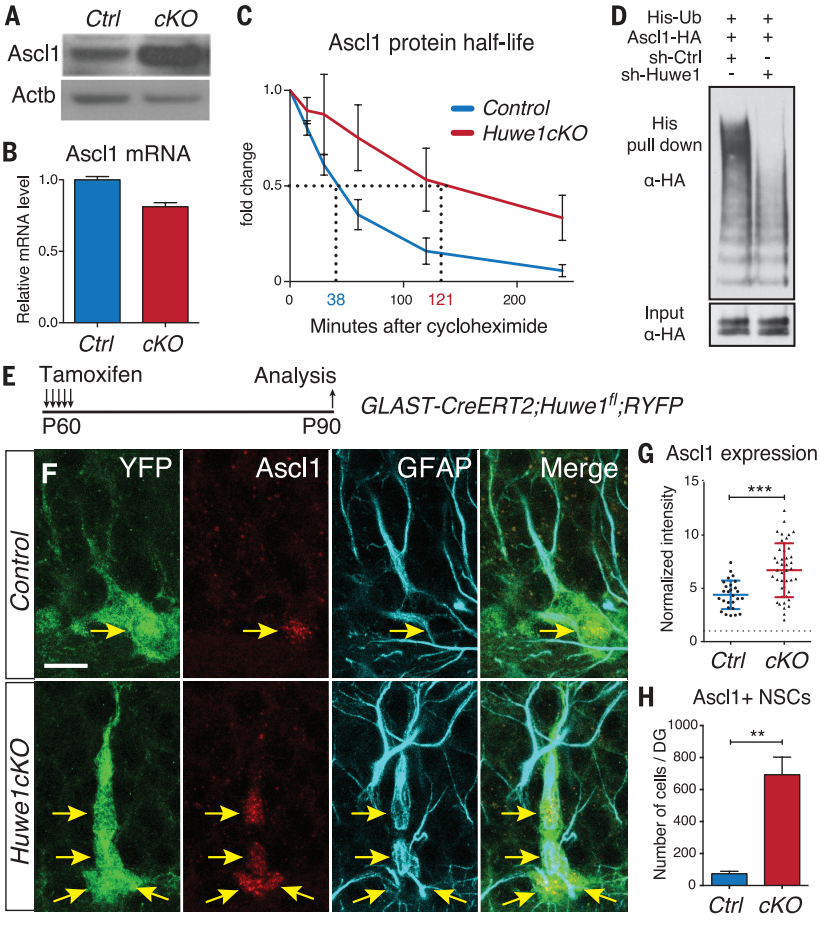



**Fig. 1. *Huwe1* controls Ascl1 stability in adult hippocampal stem cells.** (**A** and **B**) *Huwe1* inactivation in embryonic telencephalon-derived cultured NSCs increases levels of Ascl1 protein [(A) Western blot; Actin B (Actb) is used as loading control] but not of *Ascl1* mRNA [(B) quantitative polymerase chain reaction analysis of empty (Ctrl) or Cre recombinase (cKO)–expressing adenovirus-transduced cells]. (**C**) Cells were treated with cycloheximide so as to stop protein synthesis for different times and processed for Western blot to determine Ascl1 half-life; $n = 4$ independent experiments. (**D**) Ubiquitynilated Ascl1 (top) and total Ascl1 levels (bottom) were determined by immunoblotting with an antibody to hemagglutinin (HA) after transfection with *HA-Ascl1* and control or *Huwe1* shRNA. (**E** to **H**) *Huwe1* was inactivated in adult DG NSCs by means of five injections of tamoxifen at P60 followed by analysis at P90. (F) Scale bar, 10 μm. The number of Ascl1-positive NSCs, identified by their position in the subgranular zone and the presence of a GFAP$^+$ radial process [(F) and (G)], and the intensity of Ascl1 immunolabeling per cell (H), were quantified; (H) $n = 4$ mice per condition. (G) $n = 27$ Ascl1-positive cells from four mice (control) and 42 Ascl1-positive cells from four mice (*Huwe1cKO*). Yellow arrows in (F) point to Ascl1-positive cells.

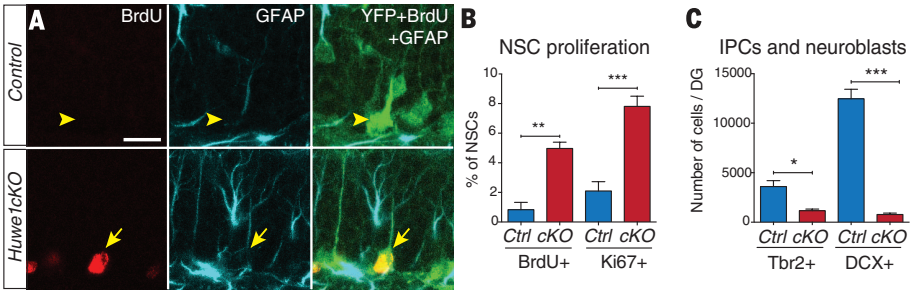
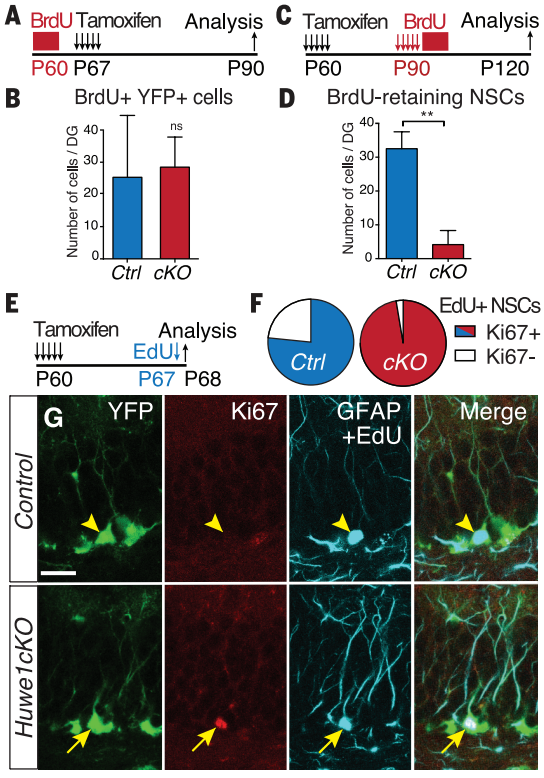



**Fig. 2. *Huwe1* inactivation promotes hippocampal stem cell proliferation and blocks progenitor differentiation.** (**A** and **B**) Hippocampal stem cell proliferation was assessed by means of Ki67 staining (B) and BrdU incorporation after a 2-hour pulse [(A) and (B)]. The total number of NSCs remained the same (fig. S6B). Yellow arrowheads point to BrdU-negative NSCs, and yellow arrows point to BrdU-positive NSCs. (A) Scale bar, 20 μm; $n = 3$ mice (BrdU) and 6 mice (Ki67) per condition. (**C**) The generation of neuronal precursors was assessed by counting the numbers of Tbr2-positive intermediate progenitors and of DCX-positive neuroblasts; $n = 3$ mice per condition.



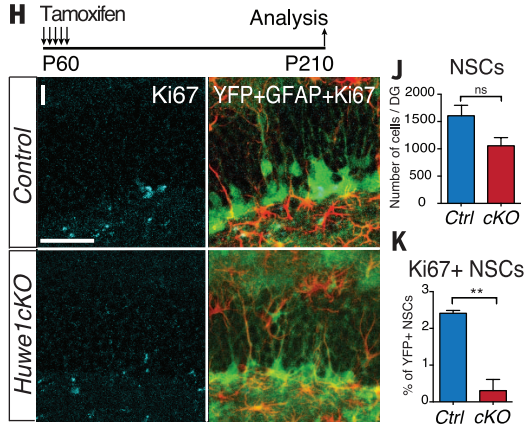




**Fig. 3. Adult hippocampal stem cells fail to return to quiescence in *Huwe1cKO* mice.** (**A** and **B**) Mice received BrdU in the drinking water for 5 days, followed by *Huwe1* inactivation. Analysis was performed 3 weeks later; $n$ = 3 (control) and 6 (*Huwe1cKO*) mice. An additional BrdU-retention experiment is shown in fig. S9, D to F. (**C** and **D**) BrdU was administered after *Huwe1* inactivation, when more *Huwe1cKO* NSCs than control NSCs proliferate (Fig. 2B). Analysis was performed 3 weeks later; $n$ = 5 (control) and 3 (*Huwe1cKO*) mice. (**E** to **G**) EdU was injected 24 hours before analysis. EdU$^+$ Ki67$^+$ cells continue proliferating, whereas EdU$^+$ Ki67$^-$ cells have exited the cell cycle. No EdU$^+$ NSCs expressed the astrocytic marker S100β (fig. S11, D to I). The yellow arrowhead points to an EdU$^+$ Ki67$^-$ NSC, and the yellow arrow points to an EdU$^+$ Ki67$^+$ NSC; $n$ = 47 (control) and 38 (*Huwe1cKO*) EdU$^+$ NSCs from 6 and 5 mice, respectively. (**H** to **K**) Analysis was performed 5 months after *Huwe1* inactivation. The overall number of NSCs was not changed, but fewer NSCs proliferated in *Huwe1cKO* than control mice; $n$ = 4 mice per condition. Scale bars, 20 μm (G) and 50 μm (I).

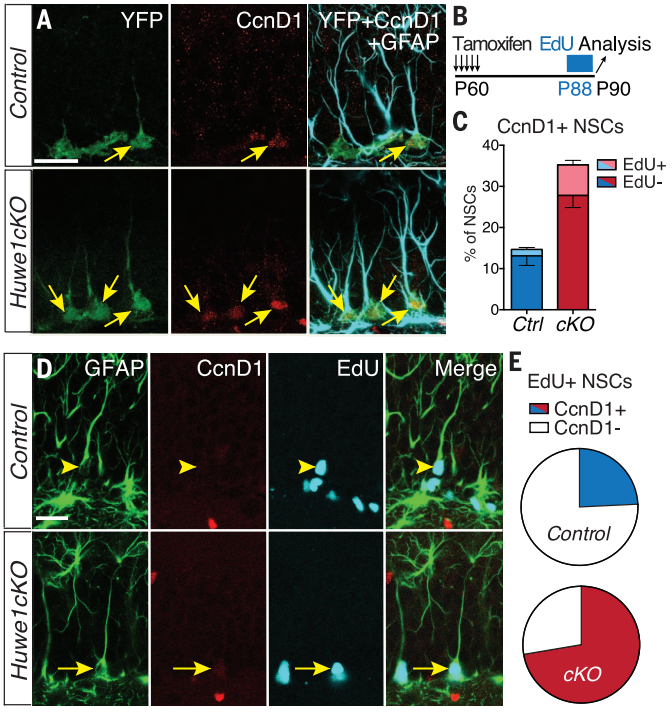



**Fig. 4. *CcnD* genes are abnormally up-regulated in *Huwe1cKO* hippocampal stem cells.** (**A** to **E**) EdU was added to the drinking water for 48 hours before the analysis to mark cells that progressed through S-phase during this period. Colabeling for EdU and CcnD1 identifies cells that have proliferated and still express CcnD1, which is required for proliferation of adult hippocampal stem cells (*17*, *18*). Pie charts in (E) show the percentage of EdU$^+$ NSCs that maintain CcnD1 expression. In (A), yellow arrows point to CcnD1$^+$ NSCs; in (D), yellow arrowheads point to EdU$^+$ CcnD1$^-$, and yellow arrows point to EdU$^+$ CcnD1$^+$ NSCs. The elevation of *CcnD1* and *CcnD2* expression was not due to an increase in proliferation of *Huwe1*-mutant NSCs [(C) and fig. S13, B and C]. (C) $n$ = 6 mice per condition and (E) $n$ = 33 (control) and 47 (Huwe1cKO) EdU$^+$ NSCs from 6 mice per condition. Scale bars, 20 μm.

## REFERENCES AND NOTES


1. T. H. Cheung, T. A. Rando, *Nat. Rev. Mol. Cell Biol.* **14**, 329–340 (2013).
2. W. Deng, J. B. Aimone, F. H. Gage, *Nat. Rev. Neurosci.* **11**, 339–350 (2010).
3. B. R. Miller, R. Hen, *Curr. Opin. Neurobiol.* **30C**, 51–58 (2014).
4. J. Andersen *et al.*, *Neuron* **83**, 1085–1097 (2014).
5. G. Froyen *et al.*, *Am. J. Hum. Genet.* **82**, 432–443 (2008).
6. S. E. McCarthy *et al.*, *Mol. Psychiatry* **19**, 652–658 (2014).
7. X. Zhao *et al.*, *Nat. Cell Biol.* **10**, 643–653 (2008).
8. Q. Zhong, W. Gao, F. Du, X. Wang, *Cell* **121**, 1085–1095 (2005).
9. J. R. Hall *et al.*, *Mol. Biol. Cell* **18**, 3340–3350 (2007).
10. J. Zhang *et al.*, *Genes Dev.* **25**, 2610–2618 (2011).
11. T. Mori *et al.*, *Glia* **54**, 21–34 (2006).
12. V. V. Terskikh, A. V. Vasiliev, E. A. Vorotelyak, *Stem Cell Rev.* **8**, 414–425 (2012).
13. M. A. Bonaguidi *et al.*, *Cell* **145**, 1142–1155 (2011).
14. D. S. Castro *et al.*, *Genes Dev.* **25**, 930–945 (2011).
15. L. C. Cheng, M. Tavazoie, F. Doetsch, *Neuron* **46**, 363–367 (2005).
16. J. M. Encinas *et al.*, *Cell Stem Cell* **8**, 566–579 (2011).
17. A. Kowalczyk *et al.*, *J. Cell Biol.* **167**, 209–213 (2004).
18. A. Ansorg, O. W. Witte, A. Urbach, *BMC Neurosci.* **13**, 46 (2012).


## ACKNOWLEDGMENTS

We gratefully acknowledge L. Chen, A. Achimastou, and E. Lavedeze for technical support; M. del Mar Masdeu, R. Subramaniams, and J. Mor for managing the mouse colony; A. Iavarone and M. Götz for providing transgenic mice; and I. Crespo, R. Kageyama, and members of the Guillemot laboratory for discussions. The study was conceived and the manuscript was written by N.U. and F.G. Most of the work was performed by N.U.; D.L.C.B., J.A.A.D., and C.H. performed the mass spectrometry analysis; J.A. helped with experiments; and A.F. and O.A. analyzed the polyubiquitination of Ascl1. N.U. was supported by fellowships from the Medical Research Council (MRC) and the Francis Crick Institute, D.L.C.B. was supported by fellowships from the

Federation of European Biochemical Societies and the Francis Crick Institute. J.A.A.D. was financed by the Netherlands Organisation for Scientific Research (NWO) project (184.032.201), and A.F. was financed by the Fondation de France. This work was supported by grants from the Wellcome Trust (106187/Z/14/Z), the Biotechnology and Biological Sciences Research Council (BB/K005316/1), the MRC (U117570528), and the Francis Crick Institute (BZ10089) to F.G. The authors declare no conflicts of interest. Additional data is available in the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/353/6296/292/suppl/DC1
Materials and Methods
Figs. S1 to S14
Table S1
References (*19*–*35*)

15 February 2016; accepted 21 June 2016
10.1126/science.aaf4802

STRUCTURAL BIOLOGY

# Structural basis for integration of GluD receptors within synaptic organizer complexes

**Jonathan Elegheert,[1] Wataru Kakegawa,[2] Jordan E. Clay,[1] Natalie F. Shanks,[3] Ester Behiels,[1] Keiko Matsuda,[2] Kazuhisa Kohda,[2] Eriko Miura,[2] Maxim Rossmann,[4] Nikolaos Mitakidis,[1] Junko Motohashi,[2] Veronica T. Chang,[1] Christian Siebold,[1] Ingo H. Greger,[4] Terunaga Nakagawa,[3]* Michisuke Yuzaki,[2]* A. Radu Aricescu[1]***

**Ionotropic glutamate receptor (iGluR) family members are integrated into supramolecular complexes that modulate their location and function at excitatory synapses. However, a lack of structural information beyond isolated receptors or fragments thereof currently limits the mechanistic understanding of physiological iGluR signaling. Here, we report structural and functional analyses of the prototypical molecular bridge linking postsynaptic iGluR δ2 (GluD2) and presynaptic β-neurexin 1 (β-NRX1) via Cbln1, a C1q-like synaptic organizer. We show how Cbln1 hexamers "anchor" GluD2 amino-terminal domain dimers to monomeric β-NRX1. This arrangement promotes synaptogenesis and is essential for D-serine–dependent GluD2 signaling in vivo, which underlies long-term depression of cerebellar parallel fiber–Purkinje cell (PF-PC) synapses and motor coordination in developing mice. These results lead to a model where protein and small-molecule ligands synergistically control synaptic iGluR function.**

Excitatory neurotransmission in the vertebrate central nervous system is largely mediated by the ionotropic glutamate receptor (iGluR) family members, classified as AMPA (GluA1-4), *N*-methyl-D-aspartate (NMDA) (GluN1, GluN2A-D, GluN3A-B), kainate (GluK1-5), and delta (GluD1-2) subtypes (*1*). All iGluRs are assembled from four modular subunits, each displaying extracellular amino-terminal and ligand-binding domains (ATD and LBD), a transmembrane domain (TMD) lining a central ion channel pore, and a cytoplasmic carboxy-terminal domain (CTD) (*2*–*6*). Binding of agonist molecules to the LBDs of GluA, GluN, and GluK receptors (typically glutamate, but also glycine or D-serine (D-Ser) for GluN subtypes) drives opening of the cation-conductive ion channel and neuronal membrane depolarization (*1*). Furthermore, all iGluRs appear to initiate postsynaptic signaling through nonionotropic mechanisms, a process better understood for NMDA and delta receptor subtypes (*7*–*15*).

In addition to their signaling function, mediated by excitatory amino acids, iGluRs are implicated in synaptogenesis (*16*–*20*). The membrane-distal position of iGluR ATDs makes them readily accessible to other proteins populating the synaptic cleft. In this context, the GluD subfamily of iGluRs are the best studied. Cbln1, a soluble synaptic organizer molecule belonging to the C1q–tumor necrosis factor α (C1q-TNFα) superfamily (*21*), directly binds the GluD1 and GluD2 ATDs (*16*, *18*, *22*, *23*). Cbln1 also interacts with presynaptic membrane-tethered neurexins (NRXs) (*18*) and establishes "molecular bridges" that span the cleft to facilitate bi-directional synaptic differentiation (*16*, *18*). Despite their importance, the architecture of supramolecular GluD assemblies and the mechanisms by which they support integration of the dual synaptogenic and metabotropic signaling functions have remained unknown. We sought to address these questions by structurally investigating the β-NRX1–Cbln1–GluD2 transsynaptic triad.

As a first step toward solving a Cbln-GluD complex, we solved high-resolution crystal structures of (i) free human GluD2 ($GluD2_{ATD}$, 1.75 Å) and mouse GluD1 ($GluD1_{ATD}$, 2.30 Å) ATD dimers and (ii) free, nearly full-length human Cbln1 ($Cbln1^{\Delta VRSG}$, 2.80 and 7.00 Å) and the obligate Cbln1 globular domain trimer (with a C1q-like fold, $Cbln1_{C1q}$, 2.35 Å) (figs. S1 to S4 and table S1) (*24*). Using surface plasmon resonance (SPR), we found that avidity governs the full-length Cbln1-GluD2 ($Cbln1_{FL}$-$GluD2_{FL}$) complex formation, which results in a nanomolar-range apparent affinity ($K_{D,app}$ of ~125 nM) (fig. S5) (*24*). Molecular dissection of the complex into individual components revealed that the Cbln1 C1q-like trimer is the minimal unit needed for interaction with $GluD2_{ATD}$, with an affinity in the high micromolar range (figs. S5 and S6) (*24*). Attempts at cocrystallizing a Cbln1-$GluD2_{ATD}$ complex were hampered by the propensity of each component to crystallize separately. We consequently designed a construct that combines a fused $Cbln1_{C1q}$ trimer with $GluD2_{ATD}$ into one continuous polypeptide chain, linked by a 30-residue, flexible Gly-Gly-Ser [$(G_2S)_{10}$] spacer (Fig. 1A and figs. S7 and S8) (*24*, *25*). We validated the mass and monodispersity of the $Cbln1_{C1q}$-$GluD2_{ATD}$ chimera using multiangle light scattering (MALS) and negative-staining single-particle electron microscopy (EM) (fig. S8) (*24*).

We determined the crystal structure of the $Cbln1_{C1q}$-$GluD2_{ATD}$ complex at 3.10 Å (Fig. 1B, fig. S8, and table S1). $Cbln1_{C1q}$ sits on top of the membrane-distal face of the $GluD2_{ATD}$ in an inward tilted orientation and breaks its threefold symmetry to engage an ATD monomer with a total buried surface area (BSA) of 873 $Å^2$ (Fig. 1C). The arrangement of both $Cbln1_{C1q}$ trimers suggests the position of the putative Cbln1 N-terminal "cysteine-rich region" (CRR, not present in the chimeric construct) that links two $Cbln1_{C1q}$ trimers into the hexameric $Cbln1_{FL}$ (Fig. 1B and fig. S4) (*26*). The distance between the calculated centers of mass of the trimers is 70 Å, in good agreement with the corresponding distances in two crystal structures of free $Cbln1^{\Delta VRSG}$ (77 and 79 Å, respectively) (fig. S9). The tilted versus linear arrangement of the C1q trimers in the three structures suggests their intrinsic hinge movement relative to the CRR (fig. S9), consistent with the single-particle EM analysis (Fig. 1D and fig. S10). The $Cbln1_{C1q}$-$GluD2_{ATD}$ structure, however, indicates that binding of GluD2 constrains the $Cbln1_{C1q}$ domain orientation. Single-particle EM class averages of free $Cbln1_{FL}$ suggested that the CRR is a globular structure linking two Cbln1 C1q trimers (Fig. 1D and fig. S10) (*24*).

In the $Cbln1_{C1q}$-$GluD2_{ATD}$ crystal lattice, $GluD2_{ATD}$ forms the same N-shaped tetrameric "AB-CD" arrangement (Fig. 2A) previously observed in the full-length GluA2 and GluK2 structures (*5*, *27*) and in structures of isolated GluA2 (*28*, *29*) and GluK6 (*30*) ATDs. Cα-atom superposition of

[1]Division of Structural Biology, Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. [2]Department of Physiology, Keio University School of Medicine, 35 Shinanomachi, Shinjuku-ku, Tokyo 160-8582, Japan. [3]Department of Molecular Physiology and Biophysics, School of Medicine, Vanderbilt University, Nashville, TN 37232-0615, USA. [4]Neurobiology Division, MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, UK.
*Corresponding author. Email: radu@strubi.ox.ac.uk (A.R.A); myuzaki@a5.keio.jp (M.Y.)

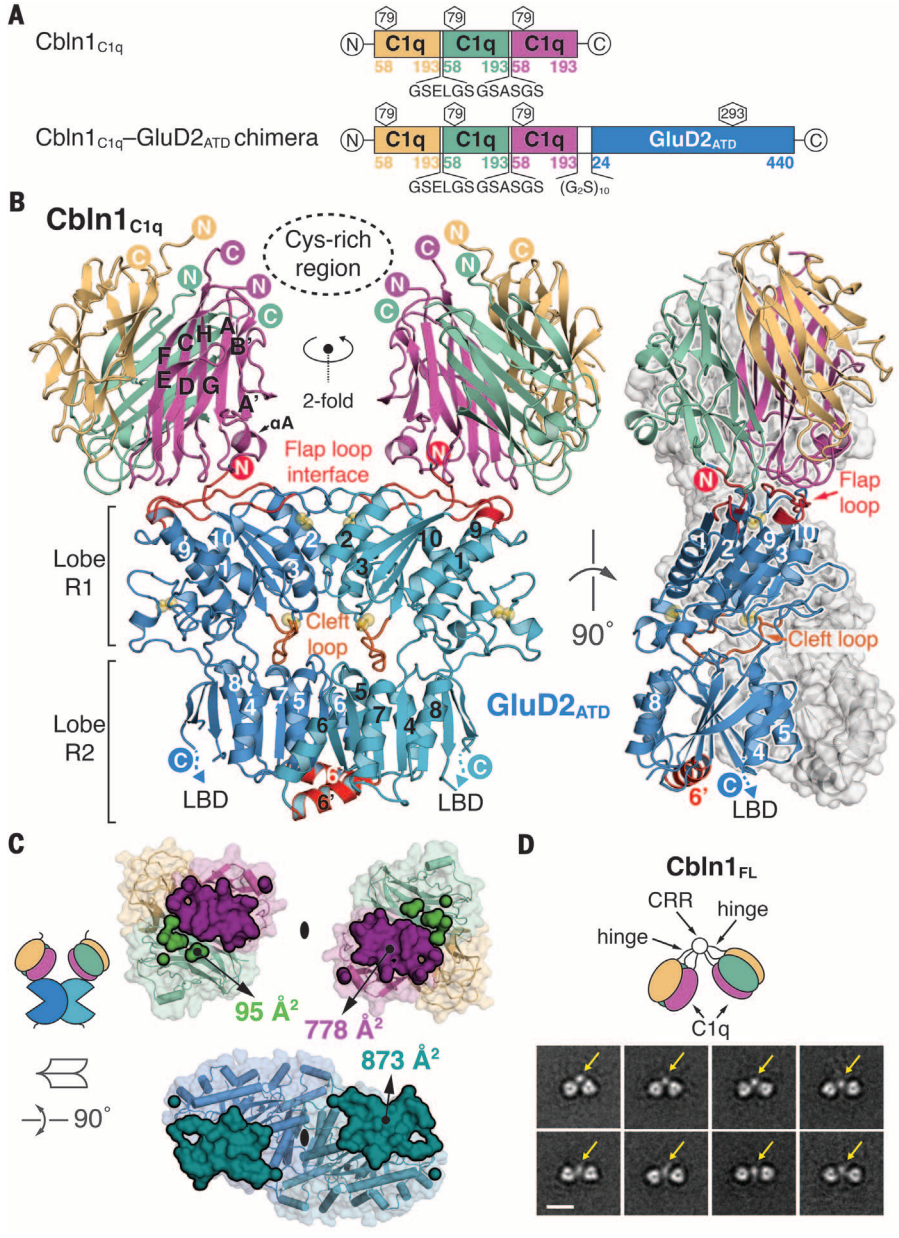



**Fig. 1. Architecture of the Cbln1$_{C1q}$-GluD2$_{ATD}$ binary complex.** (**A**) Schematic representation of the chimeric Cbln1$_{C1q}$ and Cbln1$_{C1q}$-GluD2$_{ATD}$ constructs. (**B**) "Front" and "side" view of the Cbln1$_{C1q}$-GluD2$_{ATD}$ complex. The inward tilted orientation of the Cbln1 C1q domains suggests the position of the Cbln1 CRR, as visible in EM class averages and represented by a dotted ellipse here. GluD2 α helix 6′ and flap and cleft loops are highlighted. Disulfide bridges are shown as yellow spheres. (**C**) Symmetry-breaking in the Cbln1$_{C1q}$ interface. The total buried interaction surface is shown in a 90° rotated open-book view. (**D**) Selected negative-stain EM class averages of Cbln1$_{FL}$ illustrate its dimer-of-trimers arrangement. Yellow arrows indicate the suggested position of the CRR that links both C1q trimers. Scale bar, 10 nm.

GluA2$_{CRYST}$ (*5*) and GluD2 ATD tetramers yields a root mean square deviation (RMSD) of 5.6 Å over 635 positions when aligning B-D dimers, and 7.3 Å over 695 positions when aligning A-C dimers (Fig. 2B). This AB-CD ATD arrangement is, as in GluA2$_{CRYST}$, stabilized by two-fold symmetrical contacts between the D-B monomers (666 Å$^2$ BSA) and consists of numerous putative salt bridges and hydrogen bonds, as well as potentially two calcium atoms (fig. S11) (*24*). Alignment of the Cbln1$_{C1q}$-GluD2$_{ATD}$ complex to the full GluA2$_{CRYST}$, using the B-D ATD dimer as reference, illustrates how one Cbln1 hexamer, binding to each ATD dimer, extends the outward-tapering, vertical Y-shape of the iGluR. The height of this arrangement (~17 nm) fits within the typical width of the PF-PC synaptic cleft (~20 nm) (Fig. 2C).

We then investigated how Cbln1 links GluD2 to presynaptic β-NRX1. The extracellular part of the single-span membrane-tethered β-NRX1 consists of a single LNS (laminin, neurexin, sex hormone–binding globulin) domain, LNS6, of known structure (*31*). Insertion of the 30-residue "spliced sequence #4" (SS4) into β-NRX1$_{LNS6}$ via alternative splicing, to yield β-NRX1(+4), is required for Cbln1 binding (*18*). Using isothermal titration calorimetry (ITC), we measured that one hexameric Cbln1 and one monomeric β-NRX1(+4) bind with high affinity ($K_D$ = 43.5 ± 4.4 nM) and that the Cbln1 CRR and β-NRX1 SS4 mediate all binding (fig. S12). Notably, the 1-to-1 stoichiometry indicates that the two-fold Cbln1 symmetry is broken in β-NRX1(+4)–Cbln1$_{FL}$. Single-particle negative-stain EM class averages of β-NRX1(+4)–Cbln1$_{FL}$ confirm this stoichiometry and confirm that the Cbln1 CRR is the β-NRX1(+4) binding platform (Fig. 2D and fig. S10) (*24*). Together, our structural analyses allow us to propose an overall model for the β-NRX1–Cbln1–GluD2 triad that features symmetry mismatches in both binary complex interfaces (Fig. 2E).

Cbln1$_{C1q}$ binds GluD2$_{ATD}$ at the membrane-distal ends of α helices 1, 2, 3, 9 and 10 of the R1 lobe and at the "flap" loop, an extended, structurally conserved segment that links α helices 9 and 10 and folds back onto the top of the ATD. Cbln1 interface residues are contributed by loops *AA′* (Ser$^{69}$-Ser$^{75}$), *CD* (Tyr$^{122}$-Thr$^{126}$), *EF* (Gly$^{144}$-Arg$^{150}$), and *GH* (Gly$^{174}$-Lys$^{181}$) from one C1q subunit (Fig. 3A). Cbln1 loops *EF* and *CD* rearrange upon binding the relatively flat top of GluD2$_{ATD}$, which remains largely unchanged between free and Cbln1-bound states (overall Cα-atom RMSD of 0.5 Å over 336 positions) (Fig. 3A).

Cbln1 Tyr$^{122}$ forms the central interaction hotspot; it is oriented by GluD2 Glu$^{61}$ and Thr$^{60}$ and buried in a small hydrophobic pocket formed by GluD2 Ile$^{26}$, Ile$^{27}$, Leu$^{58}$, Leu$^{342}$, and Trp$^{347}$ (fig. S13). At the periphery, Asp$^{24}$, the first residue of the mature GluD2 chain, is locked into an extended putative hydrogen-bonding network by Cbln1 residues Thr$^{70}$, Asn$^{71}$, and His$^{72}$ of loop *AA′*, and Lys$^{181}$ of loop *GH*. Cbln1 Arg$^{150}$ and Asp$^{147}$ on loop *EF* engage in putative charged interactions with GluD2 Glu$^{61}$ and Arg$^{345}$, respectively (fig. S13). Sequence conservation analysis indicates that the flap loop interface residues of GluD1/2$_{ATD}$ and the loop-based interface residues in the Cbln family (Cbln1 to Cbln4) are highly conserved in vertebrates (fig. S14).

We performed single-position alanine-scanning mutagenesis on 15 contiguous GluD2 interface residues to validate the Cbln1$_{C1q}$-GluD2$_{ATD}$ binding mode. These mutations containing an amino acid substitution (*25*)—D24A, S25A, I26A, E61A, L342A, R345A, H348A, or S352A—cluster into the center of the observed interaction interface, and they showed reduced binding to Cbln1 in an avidity-enhanced SPR setup. Seven peripheral mutations maintained binding: T60A, E343A, D344A, K346A, S349A, M350A, and Q364A (Fig. 3, B and C, and fig. S15). Furthermore, the Cbln1$_{FL}$, Cbln1$^{C34S,C38S}$ (lacking the CRR), and Cbln1$_{C1q}$ variants containing the Y122A, R124A, and D147A interface mutations lost all binding to GluD2$_{ATD}$ (fig. S16).

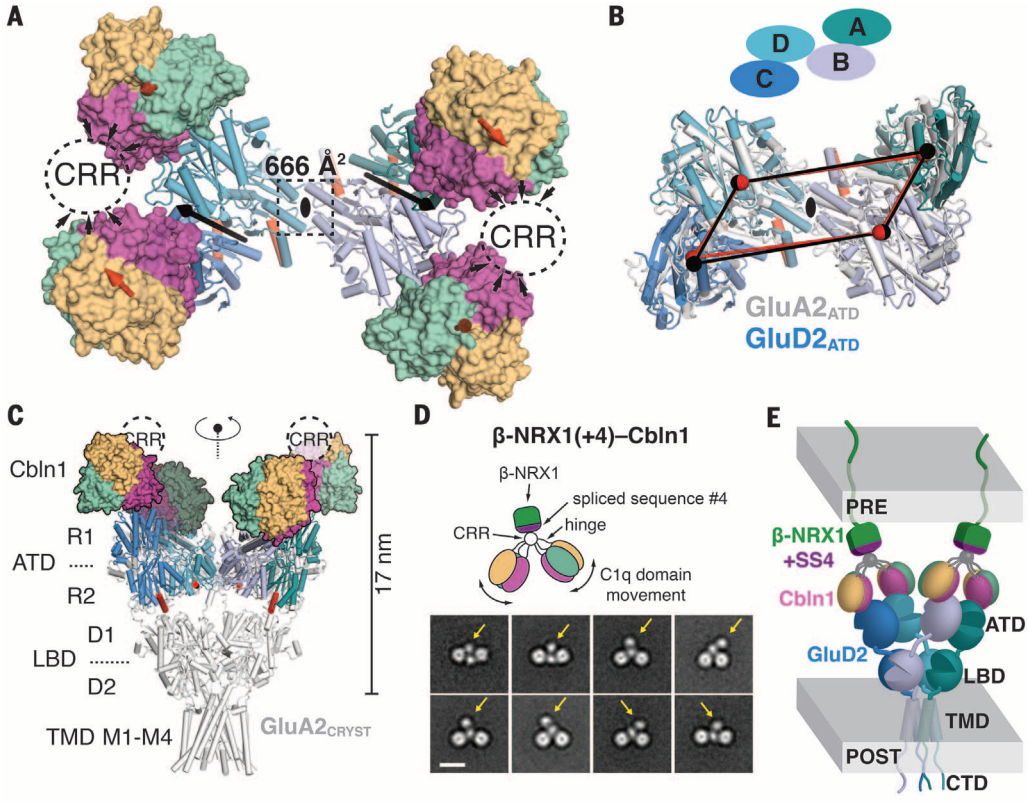



**Fig. 2. Quaternary structure of the β-NRX1–Cbln1–GluD2 complex.** (**A**) "Top" view of the full $Cbln1_{C1q}$-$GluD2_{ATD}$ dimer-of-dimers complex. The black ellipse, black arrows, and red arrows indicate the overall two-fold symmetry axis, the two-fold symmetry axes in the $GluD2_{ATD}$ dimers, and the three-fold symmetry axes in the $Cbln1_{C1q}$ trimers, respectively. The suggested position of the Cbln1 CRR is marked with dashed ovals. (**B**) Superposition of the GluD2 and GluA2 (PDB 3KG2) (*5*) N-shaped ATD layers using the B-D ATD dimers [view equivalent to (A)]. Centers of mass of $GluD2_{ATD}$ (black spheres) and $GluA2_{ATD}$ (red spheres) are connected to highlight overall similarity. (**C**) View along the overall two-fold axis of the $Cbln1_{C1q}$-$GluD2_{ATD}$ complex aligned to Y-shaped $GluA2_{CRYST}$ using the B-D ATD dimers. (**D**) Selected negative-stain EM class averages of the β-NRX1(+4)–$Cbln1_{FL}$ complex. Yellow arrows indicate the suggested position of β-NRX1(+4). Scale bar, 10 nm. (**E**) Model of the synapse-spanning β-NRX1(+4)–Cbln1–GluD2 complex.

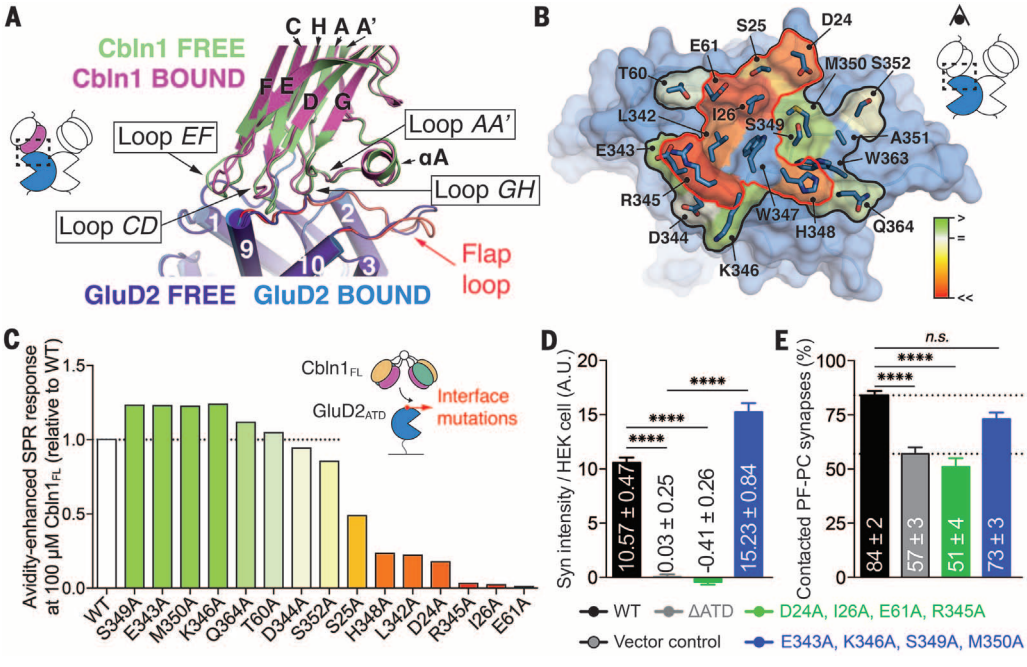



**Fig. 3. Details and structure-guided mutagenesis of the Cbln1-GluD2 interface.** (**A**) Superposition of free and bound $Cbln1_{C1q}$ and $GluD2_{ATD}$. (**B**) GluD2 alanine-scanning mutagenesis; SPR response levels are color-annotated as a heat map onto the GluD2 ATD structure. The mutated interface is outlined in black, and the interaction hotspot is outlined in red. (**C**) GluD2 alanine-scanning mutagenesis using $Cbln1_{FL}$ and monomerized $GluD2_{ATD}$. The chart shows absolute SPR responses after stimulation with 100 μM Cbln1, relative to wild-type GluD2. (**D**) Quantification of hemi-synapse formation by GCs and HEK 293T cells expressing structure-guided GluD2 mutants. Syn, synaptophysin. Data represent means ± SEM. ****$P < 0.0001$ (Kruskal-Wallis and Steel-Dwass test). (**E**) Quantification of contacted synapses between PFs and PCs expressing structure-guided GluD2 mutants in *Grid2*-null (GluD2-deficient) cerebella. Data represent means ± SEM. ****$P < 0.0001$; n.s., not significant (Kruskal-Wallis and Steel-Dwass test).

The expression pattern of GluD2, specifically confined to PF-PC synapses in the cerebellar cortex (*32*), as well as the availability of well-characterized *Grid2*-null (GluD2-deficient) and *Cbln1*-null mice, offered us the opportunity to validate and mechanistically interrogate structural models in a series of functional assays, from cell culture to in vivo. Guided by the $Cbln1_{C1q}$-$GluD2_{ATD}$ structure, we designed a number of GluD2 mutants that target distinct structural features of the receptor to gauge their effect on the two well-established major functions of GluD2 signaling: (i) PF-PC synapse formation and (ii) induction of long-term depression (LTD) of synaptic transmission (*8*, *32*). $GluD2^{D24A,I26A,E61A,R345A}$ combines four Ala mutations that abolish Cbln1 binding (Fig. 3B and fig. S17). $GluD2^{E343A,K346A,S349A,M350A}$ combines four Ala mutations that maintain Cbln1 binding but are immediately adjacent to the binding hotspot (Fig. 3B and fig. S17). $GluD2^{F76D}$ contains monomerized ATDs (fig. S6); this mutation was

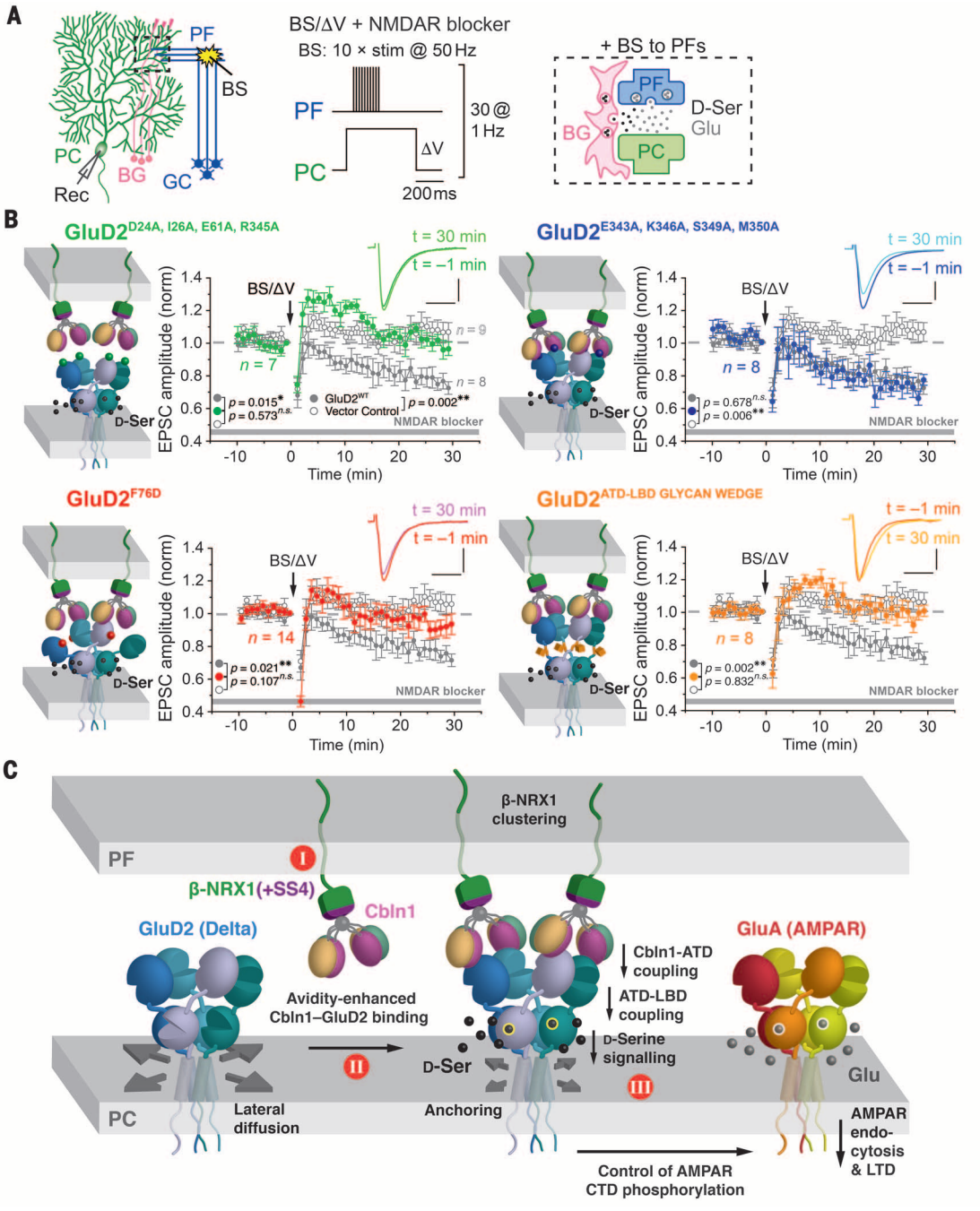



**Fig. 4. Structure-guided GluD2 mutants affect cerebellar synaptic plasticity.** (**A**) Setup to induce LTD in immature PC dendrites. BG, Bergmann glia; Rec, recording electrode; BS, burst PF stimulation; ΔV, direct PC depolarization; NMDAR, NMDA receptor. (**B**) Averaged LTD data from immature *Grid2*-null (GluD2-deficient) PCs expressing GFP + GluD2 variants, after BS/ΔV (arrow indicates). (Insets) PF-EPSCs at $t$ = –1 min and $t$ = 30 min time points relative to BS/ΔV application. Data represent means ± SEM. $^{**}P < 0.01$; $^{*}P < 0.05$; n.s., not significant (Kruskal-Wallis and Steel-Dwass test). NMDA receptor blockers are 100 μM D-AP5 plus 25 μM MK801. (**C**) Proposed key events leading to signal transmission in the β-NRX1(+4)–Cbln1–GluD2 triad. (I) β-NRX1–Cbln1 is a presynaptic anchor for GluD2. (II) Transsynaptic complex formation. GluD2 allows binding of two β-NRX1(+4)–Cbln1 complexes and is shown at full occupancy. (III) GluD2, Cbln1, and D-Ser cooperatively induce postsynaptic LTD.

specifically designed to disrupt the N-shaped ATD layer, native $Cbln1_{FL}$-$GluD2_{ATD}$–binding geometry and overall Y-shape of GluD2 (Fig. 2C), while maintaining the $Cbln1_{Clq}$-binding platform (Fig. 3A). Finally, $GluD2^{ATD\text{-}LBD\_GLYCAN\_WEDGE}$ contains a 10-residue glycosylated linker [ELSNGTDGAS in single-letter amino acid code (*25*)] inserted between the ATD and LBD layers in order to space them apart and disrupt potential mechanical ATD-LBD coupling (fig. S17) (*24*).

The effect of the glycan wedge (GW) on ATD-LBD coupling was tested by introducing the Ala654Thr mutation ["Lurcher" (Lc), located in the transmembrane (TM) helix M3] into GluD2 to render the channel constitutively open (fig. S18). Application of the LBD agonist D-Ser inhibited the Lc current in a dose-dependent manner, as expected (*33*). However, the half-maximal inhibitory concentration ($IC_{50}$) was ~1.5 times that in $GluD2^{GW/Lc}$ compared with wild-type (WT) GluD2 carrying the LC mutation ($GluD2^{WT/Lc}$), which indicated that separation of ATD and LBD by the glycan wedge impaired the ability of the LBD to induce pore closure (fig. S18).

We set up a heterologous hemi-synapse formation assay in which human embryonic kidney 293T (HEK 293T) cells expressing these mutants were cocultured with isolated wild-type mouse cerebellar granule cells (GCs) (Fig. 3D and fig. S17). Consistent with our mutagenesis data, cells expressing $GluD2^{D24A,I26A,E61A,R345A}$ accumulated no GC axon terminals, whereas cells expressing $GluD2^{E343A,K346A,S349A,M350A}$ did so normally. Cells expressing $GluD2^{F76D}$ showed intermediate GC axon terminal accumulation, which suggested that a disrupted ATD dimerization interface still allows synapse-spanning interactions to a certain extent. The $GluD2^{ATD\text{-}LBD\_GLYCAN\_WEDGE}$ mutant showed normal GC axon terminal accumulation, consistent with the notions that the $Cbln1_{Clq}$-$GluD2_{ATD}$ complex interface and the overall receptor geometry were preserved.

We used recombinant Sindbis viruses to back-express either $GluD2^{WT}$ or GluD2 mutants in the cerebellum of adult *Grid2*-null mice older than

postnatal day 30 (>P30). *Grid2*-null mice typically display a marked (~40%) reduction of PF-PC synapse number (*32*). Using immunohistochemical analysis of ultrathin sections with gold-conjugated antibodies against GluD2 (anti-GluD2), we detected that all mutants reached the PC postsynaptic density (PSD) (fig. S19). We counted the number of PF-PC synapses in the EM micrographs; the proportion of contacted synapses for GluD2$^{E343A,K346A,S349A,M350A}$ and GluD2$^{ATD\text{-}LBD_GLYCAN_WEDGE}$ was comparable to the proportion for GluD2$^{WT}$. GluD2$^{D24A,I26A,E61A,R345A}$ and GluD2$^{F76D}$, however, failed to robustly induce PF-PC contacts (Fig. 3E and fig. S19). Thus, disruption of the Cbln1-GluD2 interface and binding geometry attenuates rapid induction of PF-PC synapses.

Cerebellar LTD is caused by activity-induced endocytosis of postsynaptic AMPA receptors in PC dendrites (*34*). GluD2 signaling via the CTD is absolutely required for LTD induction, independent of PF synapse formation; LTD is impaired even when the PF synapse number is restored by reintroduction of GluD2 lacking the CTD in *Grid2*-null mice (*35*). D-Ser binds and closes the LBD of GluD2 (fig. S18) (*33*) to enhance LTD induction via the CTD of GluD2 (*8*). Application of exogenous D-Ser (200 μM for 10 min) reduced PF-evoked excitatory postsynaptic currents (PF-EPSCs) in mature wild-type, but not in *Cbln1*-null PCs, in the presence of NMDA receptor blockers to prevent coactivation of NMDA receptors (fig. S20). Endogenous D-Ser is released at PF-PC synapses from neighboring Bergmann glia by burst stimulation (BS) of PFs in immature cerebellar slices (*8*). Indeed, conjunctive BS and direct PC depolarization (BS/ΔV) (Fig. 4A) induced a robust LTD in wild-type but not in *Cbln1*-null or *Grid2*-null PCs in the presence of NMDA receptor blockers (fig. S20). These results indicate that GluD2 requires the presence of Cbln1 in order to respond to exo- or endogenous D-Ser and trigger AMPA receptor endocytosis.

We examined whether the structure-guided GluD2 mutants are able to support D-Ser–dependent LTD at PF-PC synapses. BS/ΔV induced LTD in immature *Grid2*-null PCs expressing GluD2$^{WT}$ and GluD2$^{E343A,K346A,S349A,M350A}$ but not GluD2$^{D24A,I26A,E61A,R345A}$, GluD2$^{F76D}$, or GluD2$^{ATD\text{-}LBD_GLYCAN_WEDGE}$ (Fig. 4B and fig. S21) (*24*). Thus, anchoring by Cbln1, stable ATD dimer formation, and ATD-LBD coupling are all required for GluD2 to mediate the D-Ser–dependent LTD signals.

To relate our findings to cerebellar function in vivo, we subjected immature *Grid2*-null mice virally expressing GluD2$^{WT}$ and the structure-guided GluD2 mutants to an accelerating (4 to 40 rpm. in 5 min) rotor-rod test. Conventional cerebellum-dependent learning tasks, such as eye blink-conditioning, cannot be used in immature mice, when D-Ser is still present. Motor coordination was recovered after expression of GluD2$^{WT}$, GluD2$^{E343A,K346A,S349A,M350A}$, and GluD2$^{ATD\text{-}LBD_GLYCAN_WEDGE}$ but not of GluD2$^{D24A,I26A,E61A,R345A}$ or GluD2$^{F76D}$ (fig. S22). These results reflect our PF-PC synapse formation data and suggest that the interface and geometry of the Cbln1-GluD2 complex we describe is crucial for restoration of PF-PC cerebellar circuitry and motor-related performance.

Our results provide a molecular framework for the large body of studies on the β-NRX1–Cbln1–GluD2 transsynaptic signaling system and suggest a three-step model for GluD2 signaling activation. Cbln1 is secreted from cerebellar GCs through as-yet-unidentified mechanisms and remains associated with β-NRX1 on the surface of GC axons (step I in Fig. 4C) (*36*). When cerebellar GC axons encounter PC dendritic spines, the β-NRX1–Cbln1 complex "hooks" GluD2, via avidity-enhanced Cbln1$_{C1q}$-GluD2$_{ATD}$ interactions, in a transsynaptic complex (step II in Fig. 4C). The biological importance of the binding avidity most likely lies in improving the probability for molecular recruitment of GluD2 by β-NRX1–Cbln1 at PF-PC synapses, given the weak affinity between individual Cbln1 and GluD2 domains. It is, however, unclear whether GluD2 is fully occupied at any given time point. Finally, agonist binding triggers a conformational change of the GluD2 LBD that is transmitted to the transmembrane domain to initiate downstream signaling (*9*), which results in AMPA receptor endocytosis and LTD (*8*, *9*) (step III in Fig. 4C).

Recent structures of isolated full-length GluA and GluK receptors in different functional states have established that iGluR gating is accompanied by complex ATD-LBD relative motions (*27*, *37*, *38*). Furthermore, desensitization of GluA receptors results in a marked conformational variability in the ATD layer (*4*, *27*, *37*). Similar motions might, in principle, be possible in GluD receptors, considering their overall structural and sequence homology to GluA and GluK family members. However, we propose that, in a transsynaptic context, anchoring of GluD to the β-NRX1(+4)–Cbln1 complex will limit or prevent large-scale motions of the ATD layer. As a result, the force generated by LBD closure, driving the overall receptor contraction, is likely to transfer predominantly toward the postsynaptic membrane. Because most, if not all, iGluR family members are anchored to synaptic cleft proteins via their ATDs (*16*–*20*), it is conceivable that their range of motions may also differ from those currently described in isolated receptors (*4*, *27*, *37*). Such interactions will impact on both iGluR location and conformation in response to agonist binding. We propose that the concept illustrated here, where small molecule and protein ligands cooperate in order to modulate GluD2 signaling, is likely to be more generally applicable to neurotransmitter receptors.

## REFERENCES AND NOTES


1. S. F. Traynelis *et al.*, *Pharmacol. Rev.* **62**, 405–496 (2010).
2. E. Karakas, H. Furukawa, *Science* **344**, 992–997 (2014).
3. C. H. Lee *et al.*, *Nature* **511**, 191–197 (2014).
4. T. Nakagawa, Y. Cheng, E. Ramm, M. Sheng, T. Walz, *Nature* **433**, 545–549 (2005).
5. A. I. Sobolevsky, M. P. Rosconi, E. Gouaux, *Nature* **462**, 745–756 (2009).
6. B. Herguedas *et al.*, *Science* **352**, aad3873 (2016).
7. W. Kakegawa, K. Kohda, M. Yuzaki, *J. Physiol.* **584**, 89–96 (2007).
8. W. Kakegawa *et al.*, *Nat. Neurosci.* **14**, 603–611 (2011).
9. K. Kohda *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **110**, E948–E957 (2013).
10. J. Aow, K. Dore, R. Malinow, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 14711–14716 (2015).
11. K. Dore, J. Aow, R. Malinow, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 14705–14710 (2015).
12. N. L. Weilinger *et al.*, *Nat. Neurosci.* **19**, 432–442 (2016).
13. C. Auger, D. Ogden, *J. Physiol.* **588**, 3063–3074 (2010).
14. T. Hayashi, H. Umemori, M. Mishina, T. Yamamoto, *Nature* **397**, 72–76 (1999).
15. R. J. Rodrigues, J. Lerma, *Wiley Interdiscip. Rev. Membr. Transp. Signal* **1**, 399–410 (2012).
16. K. Matsuda *et al.*, *Science* **328**, 363–368 (2010).
17. G. M. Sia *et al.*, *Neuron* **55**, 87–102 (2007).
18. T. Uemura *et al.*, *Cell* **141**, 1068–1079 (2010).
19. L. Saglietti *et al.*, *Neuron* **54**, 461–477 (2007).
20. K. Matsuda *et al.*, *Neuron* **90**, 752–767 (2016).
21. H. Hirai *et al.*, *Nat. Neurosci.* **8**, 1534–1541 (2005).
22. K. Ryu, M. Yokoyama, M. Yamashita, T. Hirano, *Biochem. Biophys. Res. Commun.* **417**, 157–161 (2012).
23. M. Yasumura *et al.*, *J. Neurochem.* **121**, 705–716 (2012).
24. Materials and methods are available as supplementary materials on *Science* Online.
25. Single-letter abbreviations for the amino acid residues are as follows: 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; and Y, Tyr.
26. D. Bao, Z. Pang, J. I. Morgan, *J. Neurochem.* **95**, 618–629 (2005).
27. J. R. Meyerson *et al.*, *Nature* **514**, 328–334 (2014).
28. A. Clayton *et al.*, *J. Mol. Biol.* **392**, 1125–1132 (2009).
29. R. Jin *et al.*, *EMBO J.* **28**, 1812–1823 (2009).
30. J. Kumar, P. Schuck, R. Jin, M. L. Mayer, *Nat. Struct. Mol. Biol.* **16**, 631–638 (2009).
31. J. Koehnke *et al.*, *Neuron* **67**, 61–74 (2010).
32. N. Kashiwabuchi *et al.*, *Cell* **81**, 245–252 (1995).
33. P. Naur *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 14116–14121 (2007).
34. M. Ito, K. Yamaguchi, S. Nagao, T. Yamazaki, *Prog. Brain Res.* **210**, 1–30 (2014).
35. W. Kakegawa *et al.*, *J. Neurosci.* **28**, 1460–1468 (2008).
36. A. Ito-Ishida *et al.*, *Neuron* **76**, 549–564 (2012).
37. K. L. Dürr *et al.*, *Cell* **158**, 778–792 (2014).
38. M. V. Yelshanskaya, M. Li, A. I. Sobolevsky, *Science* **345**, 1070–1074 (2014).


## ACKNOWLEDGMENTS


We thank staff at Diamond Light Source, T. Walter and K. Harlos for crystallization technical support, N. Scull for assistance with molecular biology, and E.Y. Jones and P. Miller for comments on the manuscript. This work was funded by the UK Medical Research Council (MRC) (G0700232 and L009609 to A.R.A. and MC_U105174197 to I.H.G.), the Japan Society for the Promotion of Science (15H05772 to M.Y. and 26117515, 26293042 to W.K.), the Core Research for Evolutional Science and Technology (CREST), Japan Science and Technology Agency (to M.Y.), the Takeda Science Foundation (W.K. and M.Y.), the Yamada Science Foundation (W.K.), the Human Frontier Science Program (RGP0065/2014 to M.Y. and A.R.A.), and the NIH (R01HD061543 to T.N.). The Wellcome Trust Centre for Human Genetics is supported by Wellcome Trust grant 090532/Z/09/Z. J.E. was supported by European Molecular Biology Organization (ALTF 1116-2012) and Marie-Curie (FP7-328531) fellowships, C.S. is a Cancer Research UK senior research fellow (C20724/A14414), and A.R.A. is an MRC senior research fellow (MR/L009609/1). Structure factors and coordinates of Cbln1$_{C1q}$, Cbln1$^{\Delta VRSG}$ crystal forms 1 and 2, GluD2$_{ATD}$, GluD1$_{ATD}$, and Cbln1$_{C1q}$–GluD2$_{ATD}$ are deposited in the Protein Data Bank (PDB codes 5KC5, 5KC6, 5KC7, 5KC8, 5KC9 and 5KCA, respectively).


## SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/353/6296/295/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S22
Table S1
References (*39*–*103*)


7 December 2015; accepted 17 June 2016
10.1126/science.aae0104

NEUROBIOLOGY

# Chromatin remodeling inactivates activity genes and regulates neural coding

Yue Yang,[1]* Tomoko Yamada,[1]*† Kelly K. Hill,[1,2] Martin Hemberg,[3] Naveen C. Reddy,[1] Ha Y. Cho,[1] Arden N. Guthrie,[1] Anna Oldenborg,[1] Shane A. Heiney,[4] Shogo Ohmae,[4] Javier F. Medina,[4] Timothy E. Holy,[1] Azad Bonni[1]‡

**Activity-dependent transcription influences neuronal connectivity, but the roles and mechanisms of inactivation of activity-dependent genes have remained poorly understood. Genome-wide analyses in the mouse cerebellum revealed that the nucleosome remodeling and deacetylase (NuRD) complex deposits the histone variant H2A.z at promoters of activity-dependent genes, thereby triggering their inactivation. Purification of translating messenger RNAs from synchronously developing granule neurons (Sync-TRAP) showed that conditional knockout of the core NuRD subunit Chd4 impairs inactivation of activity-dependent genes when neurons undergo dendrite pruning. Chd4 knockout or expression of NuRD-regulated activity genes impairs dendrite pruning. Imaging of behaving mice revealed hyperresponsivity of granule neurons to sensorimotor stimuli upon Chd4 knockout. Our findings define an epigenetic mechanism that inactivates activity-dependent transcription and regulates dendrite patterning and sensorimotor encoding in the brain.**

Neuronal activity influences transcription in neurons, and hence regulates neural circuits (*1*, *2*). Activity-dependent genes are often rapidly transcribed and then rapidly inactivated (*3*, *4*). However, attention has focused on the induction of transcription (*1*, *5*–*7*) rather than the biological roles and mechanisms of inactivation of activity-dependent transcription.

Epigenetic regulators, including adenosine 5′-triphosphate (ATP)–dependent chromatin-remodeling enzymes (*8*, *9*), are ideally suited to orchestrate the effects of neuronal activity on transcription globally. The ATP-dependent nucleosome remodeling and deacetylase (NuRD) complex triggers alterations of histone modifications, resulting in promoter or enhancer decommissioning and prolonged silencing of transcription (*10*–*15*).

To probe the role of the NuRD complex in dynamic regulation of transcription in the brain, we characterized the genome-wide occupancy of the core NuRD ATPase-encoding subunit, Chd4, in the mouse cerebellum. A substantial number of regions (9842) occupied by Chd4 in the cerebellum overlapped with transcription start sites (TSSs) (fig. S1A). Nearly all Chd4-bound TSSs (96%) harbored the histone modification H3K4 (histone H3 lysine 4) trimethylation (H3K4me3), which marks active and poised promoters (*16*, *17*), but not H3K27 trimethylation (H3K27me3), which marks inactive promoters (*18*) (Fig. 1A and fig. S1, B and C). Chd4 binding at H3K4me3-enriched TSSs was diminished in the cerebellum in mice in which Chd4 was conditionally deleted in granule neurons (Fig. 1, A and B, and fig. S1, D to H). Chd4 binding at H3K4me3-enriched TSSs in the cerebellum correlated tightly with acetylation of H3K9 and H3K14 (H3K9/14ac), which marks actively transcribed loci (*19*), and with gene expression (Fig. 1C and fig. S1I). Thus, Chd4 occupies the promoters of most actively transcribed genes in the mouse brain.

The NuRD complex triggers the sustained repression of only a small set of <200 genes in the cerebellum through diminution of H3K9/14ac and H3K4me3 at these promoters (*13*). We, therefore, reasoned that the NuRD complex might operate through another epigenetic mechanism to regulate the much larger set of Chd4-bound active genes. Exchange of the histone variant H2A.z is associated with regulation of transcription (*20*–*23*). H2A.z was enriched at 97% of Chd4-bound promoters in the mouse cerebellum (fig. S1J). Chd4 knockout decreased H2A.z and acetylated H2A.z enrichment at promoters with high Chd4 occupancy (Fig. 1, D and E), but not at most enhancers in the cerebellum (fig. S2).

We next intersected RNA sequencing (RNA-Seq) (*13*) and H2A.z chromatin immunoprecipitation (ChIP)–Seq analyses in the mouse cerebellum of conditional Chd4 knockout and control littermate mice. Although a small group of 121 up-regulated genes in conditional Chd4 knockout mice harbored increased H2A.z enrichment at their promoters, a much larger group of 1233 up-regulated genes displayed reduced H2A.z enrichment at TSSs (Fig. 1F and fig. S3, A and B). Gene ontology analyses revealed that genes with reduced H2A.z enrichment encoded proteins that function in intracellular signaling cascades, cell cycle control, and phosphorylation (fig. S3C). Notably, there was little or no change in H3K9/14ac, H3K4me3, and H3K27me3 or in the density of histone H3 at the promoters of these genes upon Chd4 knockout (Fig. 1G and fig. S3, D to G). Together, these data indicate that the NuRD complex triggers the deposition of H2A.z at the promoters of a large group of actively transcribed signaling genes and inactivates their expression in the brain in vivo.

The identification of an epigenetic link from the NuRD complex to H2A.z at the promoters of signaling genes led us to investigate whether the NuRD complex regulates transcription dynamically in response to neuronal activity. Expression of the activity-dependent genes c-*fos*, *nr4a1*, *dusp1*, and *nfil3* was increased in granule neurons of the rodent cerebellum upon membrane depolarization and rapidly inactivated 1 hour after cessation of membrane depolarization (Fig. 2A). Depletion of Chd4 or Mbd3, another subunit of the NuRD complex (*24*), impaired inactivation, but not reactivation, of activity genes in neurons after membrane depolarization (Fig. 2A and figs. S4 and S5, A and B). Thus, the NuRD complex appears to be required for inactivation of activity-dependent genes in neurons.

In chromatin immunoprecipitation followed by quantitative polymerase chain reaction (ChIP-qPCR) analyses, H2A.z increased at the promoters of the c-*fos*, *nr4a1*, and *dusp1* genes in granule neurons during the inactivation phase after membrane depolarization (fig. S5C). Depletion of Chd4 reduced H2A.z enrichment, but not histone H3, at the promoters of the c-*fos*, *nr4a1*, and *dusp1* genes in neurons during the inactivation phase (Fig. 2B) but not the activation phase (fig. S5D). Thus, the NuRD complex appears to specifically stimulate the loading of H2A.z at the promoters of activity-dependent genes during the inactivation phase of transcription.

Depletion of H2A.z by RNA interference (RNAi) in neurons increased expression of the c-*fos*, *nr4a1*, *dusp1*, and *nfil3* genes during the inactivation phase of activity-dependent transcription, with little or no effect during activation or reactivation (Fig. 2C and fig. S6). Thus, the NuRD complex and H2A.z are required for inactivation of activity-dependent gene expression in neurons.

We next used a rotarod procedure to induce neuronal activity in the mouse cerebellum (Fig. 2D). RNA-Seq analyses from the cerebellum of mice running on a rotarod for 1 hour compared to mice housed in a cage revealed increased transcription of activity genes (Fig. 2E), and these genes were inactivated within 1 hour after rotarod activity stopped. Chd4 knockout increased the expression of c-*fos*, *fosl2*, and *dusp1* in the cerebellum after cessation of rotarod activity but not during the activation phase of rotarod-induced transcription (Fig. 2F). Thus, the NuRD complex affects specifically the inactivation of activity genes in the brain.

We next tested whether NuRD-dependent inactivation of activity genes might regulate granule neuron connectivity in the cerebellum. We

[1]Department of Neuroscience, Washington University School of Medicine, St. Louis, MO 63110, USA. [2]MD-PhD Program, Washington University School of Medicine, St. Louis, MO 63110, USA. [3]Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK. [4]Department of Neuroscience, Baylor College of Medicine, Houston, TX 77030, USA.
*These authors contributed equally to this work. †Present address: Faculty of Medicine, University of Tsukuba, Tsukuba, Ibaraki, 305-8575, Japan. ‡Corresponding author. Email: bonni@wustl.edu

determined the stage of granule neuron maturation in vivo during which the NuRD complex regulates activity-dependent transcription. We used in vivo electroporation and translating ribosomal affinity purification (*25*) to characterize gene expression in developmentally synchronized granule neurons in vivo (Sync-TRAP) (Fig. 3A and fig. S7, A and B). Sync-TRAP followed by qPCR or by RNA-Seq analyses of the cerebellum in mice 6 days after electroporation revealed that expression of granule neuron–specific genes was enriched (fig. S7C), and led to the identification of chromatin regulators and ubiquitin ligases enriched in developing granule neurons (Fig. 3B and fig. S7D). Sync-TRAP-Seq analyses in *Chd4*$^{loxP/loxP}$ mice electroporated with the recombinase Cre revealed that 86% of significantly differentially expressed genes were up-regulated upon conditional Chd4 knockout (Fig. 3C and fig. S7E). Sync-TRAP-Seq and Sync-TRAP-qPCR analyses revealed that transcription of the activity-dependent *npas4*, *nfil3*, c-*fos*, and *nr4a1* genes, but not of granule neuron–specific genes, was increased in granule neurons depleted of Chd4 in vivo (Fig. 3, D and E). Immunohistochemical analyses confirmed that c-Fos protein was up-regulated in Chd4-depleted granule neurons in vivo (fig. S8, A and B). Sync-TRAP-qPCR analyses also revealed that depletion of H2A.z increased c-*fos* gene expression in granule neurons in vivo (fig. S8C). The NuRD complex and H2A.z thus appear to trigger the inactivation of activity-dependent genes in synchronously developing granule neurons in the mouse cerebellum.

Because the NuRD/H2A.z epigenetic link regulates activity-dependent transcription in a temporal window of dendrite morphogenesis in the cerebellum, we asked whether NuRD-dependent inactivation of genes might influence dendrite patterning and connectivity. Granule neurons labeled by in vivo electroporation undergo distinct stages of dendrite morphogenesis in a synchronized manner in vivo (Fig. 4A and fig. S9, A to C). Depletion of Chd4 increased the total length of granule neuron dendrites and the number of primary dendrites during the period of pruning, with little or no effect on the development of granule neuron dendrites during earlier stages (Fig. 4B and fig. S9, A to C). Expression of NuRD-regulated activity genes also impaired dendrite pruning in vivo (Fig. 4, C

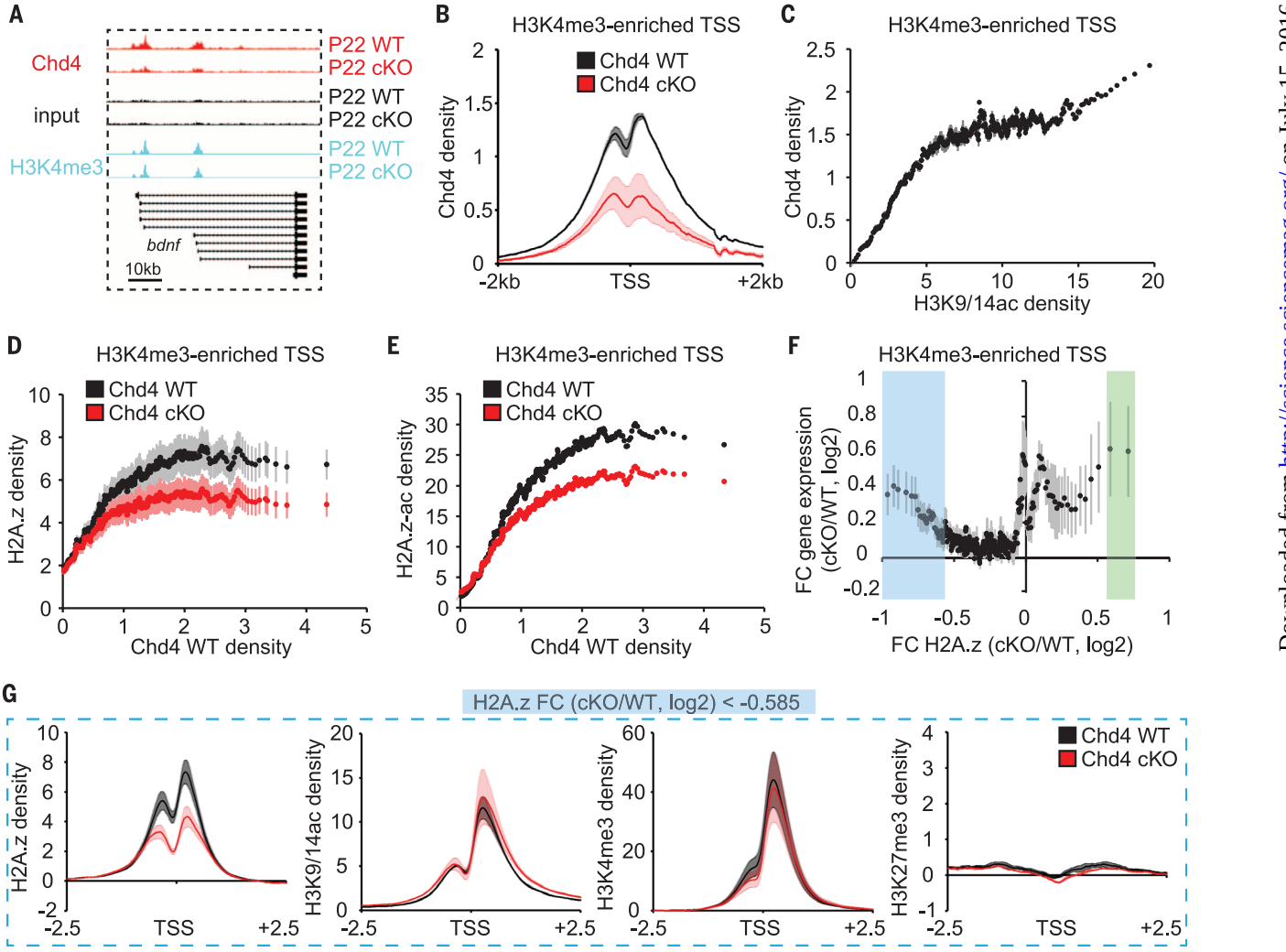



**Fig. 1. The core NuRD subunit Chd4 occupies promoters of actively transcribed genes in the cerebellum in vivo.** (**A**) University of California, Santa Cruz (UCSC) genome browser tracks at the *bdnf* locus in the cerebellum of conditional Chd4 knockout and control mice. (**B**) Location of Chd4 binding near the TSSs of H3K4me3-enriched genes in the cerebellum. In all ChIP-Seq analyses, shading denotes SE. (**C**) Comparison of Chd4 and H3K9/14ac read densities at H3K4me3-enriched genes. (**D** and **E**) Comparison of Chd4 and H2A.z [(D), $P < 0.01$, Hotelling $T^2$ test for small sample size] or acetylated H2A.z (E) read density at H3K4me3-enriched TSSs in conditional Chd4 knockout and control mice. (**F**) Comparison of the fold change in H2A.z read density and fold change in gene expression at H3K4me3-enriched TSSs in postnatal day 22 (P22) conditional Chd4 knockout and control mice. Genes with reduced H2A.z [fold change ($\log_2$) < −0.585] upon conditional Chd4 knockout are highlighted in blue, and genes with increased H2A.z (fold change ($\log_2$) > 0.585) are highlighted in green. (**G**) The profile of H2A.z, H3K9/14ac, H3K4me3, and H3K27me3 surrounding the TSSs of the group of genes indicated in (F) highlighted in blue.

and D). In contrast, upon expression of NuRD-repressed target genes not known to be regulated by activity or other cerebellum-enriched transcriptional regulators, only one increased dendrite length but not dendrite number (Fig. 4, C and D, and fig. S9D). These results indicate that NuRD-dependent inactivation of activity genes may drive granule neuron dendrite pruning in the cerebellum.

We next characterized the consequences of NuRD actions on responses of granule neurons in behaving mice. Mature granule neurons receive on average four mossy fiber inputs, which is optimal for sparse, lossless encoding of sensorimotor information (*26*). We electroporated mouse pups with a plasmid encoding the calcium indicator GCaMP6s together with an mCherry expression plasmid, and subjected mice to a motorized treadmill task (Fig. 5A and movie S1). After habituation, two-photon imaging of lobule VI in the mouse cerebellum revealed that a set of GCaMP6s-labeled granule neurons was active during locomotion (Fig. 5, B and C). Conditional Chd4 knockout triggered a robust increase in the fraction of high-fidelity treadmill-responsive granule neurons and concomitantly reduced the fraction of unresponsive granule neurons

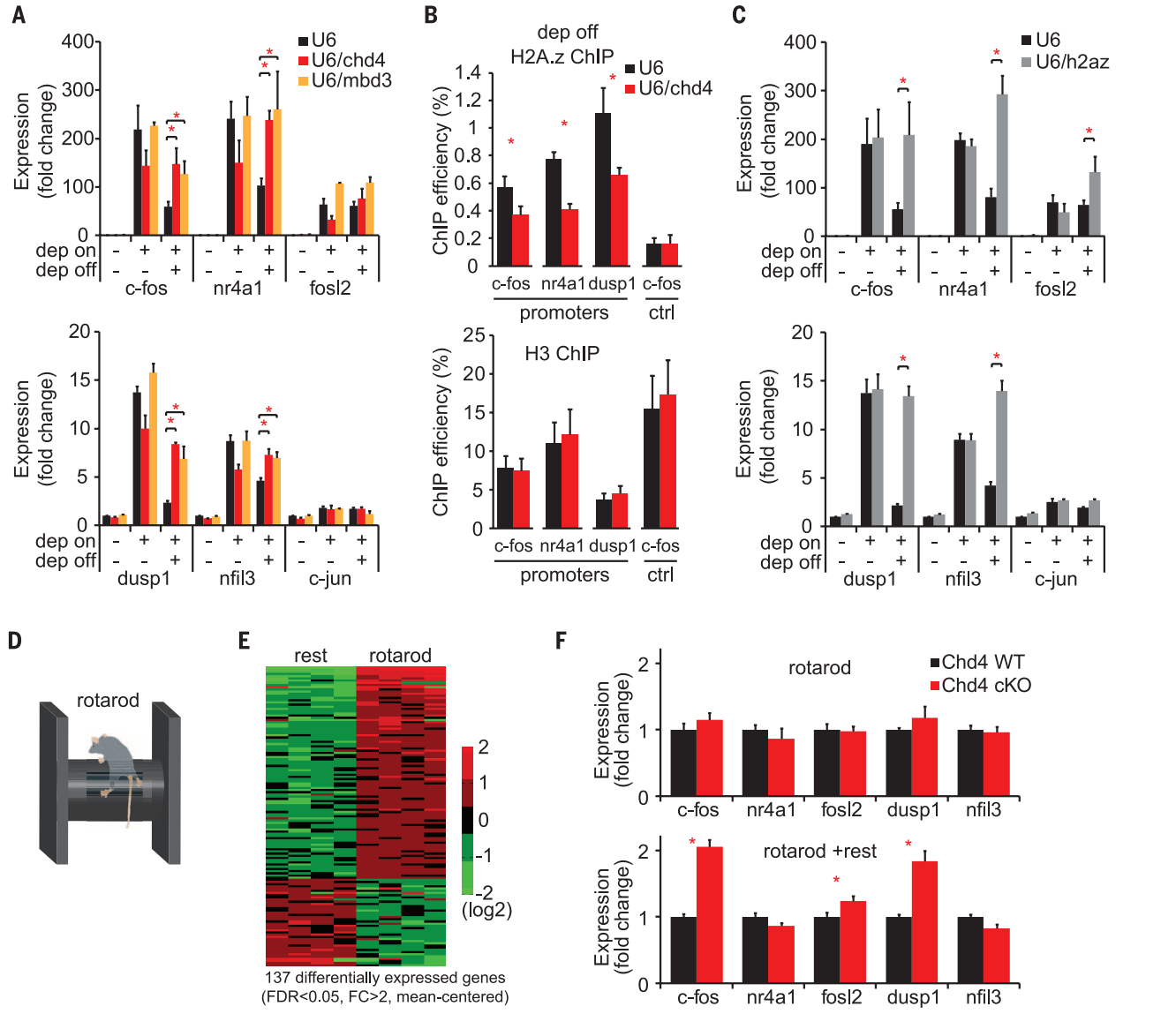



**Fig. 2. The NuRD/H2A.z chromatin-remodeling pathway inactivates activity-dependent genes in neurons.** (**A**) Granule neurons from P5 rat pups transfected with the U6/chd4, U6/mbd3, or the control U6 RNAi plasmid were depolarized (dep on) with 50 mM KCl for 1 hour and then switched back to hyperpolarizing media (dep off) for 1 hour and subjected to quantitative reverse transcription (qRT)–PCR analyses. Expression of c-*fos*, *nr4a1*, *dusp1*, and *nfil3* genes upon cessation of membrane depolarization in neurons after knockdown of the NuRD subunits Chd4 and Mbd3 (*$P < 0.05$, analysis of variance (ANOVA) followed by Fisher's protected least significant difference (PLSD) post hoc test, $n = 3$ independent experiments). (**B**) Lysates of granule neurons transfected with the U6/chd4 or control U6 RNAi plasmid and treated as in (A) were subjected to ChIP-qPCR analyses with antibody to H2A.z (top) or to histone H3 (bottom) and primers specific to the c-*fos*, *nr4a1*, and *dusp1* gene promoters or a c-*fos* control region (*$P < 0.05$, *t* test, $n = 3$ independent experiments). (**C**) Expression of c-*fos*, *nr4a1*, *fosl2*, *dusp1*, and *nfil3* genes upon cessation of membrane depolarization in granule neurons after knockdown of H2A.z (*$P < 0.05$, *t* test, $n = 3$ independent experiments). (**D** and **E**) The cerebellum of P27-P28 mice trained on the rotarod task (D) was subjected to RNA-Seq analyses. A heatmap of the expression levels of significantly differentially expressed genes (E) [false discovery rate (FDR) < 0.05, fold change > 2 for rotarod compared to control homecage, $n = 4$ mice, base 2 log-transformed mean centered]. (**F**) The cerebella of conditional Chd4 knockout or control mice trained on the rotarod task were subjected to qRT-PCR analyses (*$P < 0.05$, *t* test, $n = 6$ to 8 mice).

(Fig. 5, C and D). Thus, inhibition of the NuRD complex leads to hyperresponsivity of granule neurons to sensorimotor stimuli.

In behavior analyses, depletion of Chd4 in granule neurons impaired procedural learning, including in the accelerating rotarod and delay eyeblink conditioning assays, but had little or no effect on motor coordination as assessed in the DigiGait and open field assays (fig. S10).

Our study defines chromatin-remodeling events that inactivate activity-dependent transcription and control dendrite architecture and sensorimotor encoding in the brain (see model in Fig. 5E). Our findings suggest that inactivation of activity genes is essential for the maturation of granule neuron dendrite arbors and in the control of neural circuit activity in response to sensorimotor signals. We have uncovered the NuRD

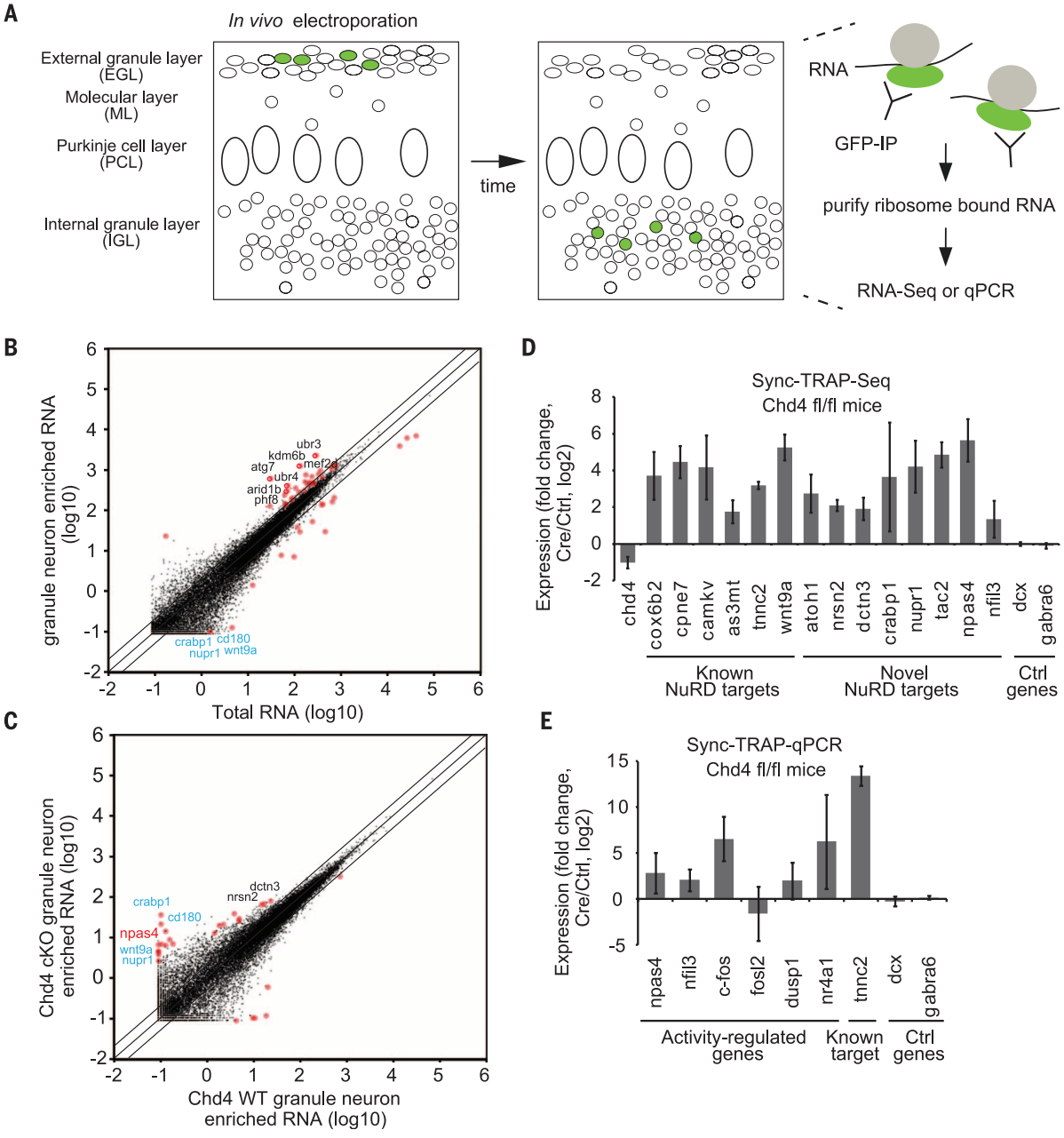



**Fig. 3. In vivo Sync-TRAP analyses reveal that the NuRD complex inactivates activity-dependent genes in synchronously developing granule neurons in the cerebellum.** (**A**) A schematic depicting the Sync-TRAP protocol. In vivo electroporation of mouse pups with the green fluorescent protein (GFP)–Rpl10a expression vector labels granule neuron precursors localized in the external granule layer (EGL). Labeled granule neurons migrate to the internal granule layer (IGL) and undergo differentiation in a synchronized manner (see fig. S9A). mRNAs bound to GFP-labeled ribosomes were profiled to characterize the in vivo gene expression program in synchronously developing granule neurons. (**B** to **E**) Sync-TRAP followed by RNA-Seq or qPCR analyses with *Chd4*$^{loxP/loxP}$ mice electroporated with the pCAG-Cre or control vector. Scatterplot of input RNA reads per kilobase of transcript per million mapped reads (RPKM) and immunoprecipitated mRNA RPKM subjected to RNA-Seq analyses (B). Scatterplot of immunoprecipitated mRNA RPKM from Cre-expressing or control granule neurons subjected to RNA-Seq analyses (C). Genes that are less abundant in control electroporated granule neurons compared to input (B) and increased upon knockout of Chd4 compared to control (C) are denoted in light blue. Diagonal lines represent 0.5-, 1-, and 2-fold changes in the geometric mean of gene expression between conditions [(B) and (C), red circles denote FDR < 0.05]. Fold changes in gene expression of NuRD targets identified by Sync-TRAP followed by RNA-Seq (D). Validation of NuRD-regulated activity-dependent genes with Sync-TRAP followed by qPCR analyses (E).

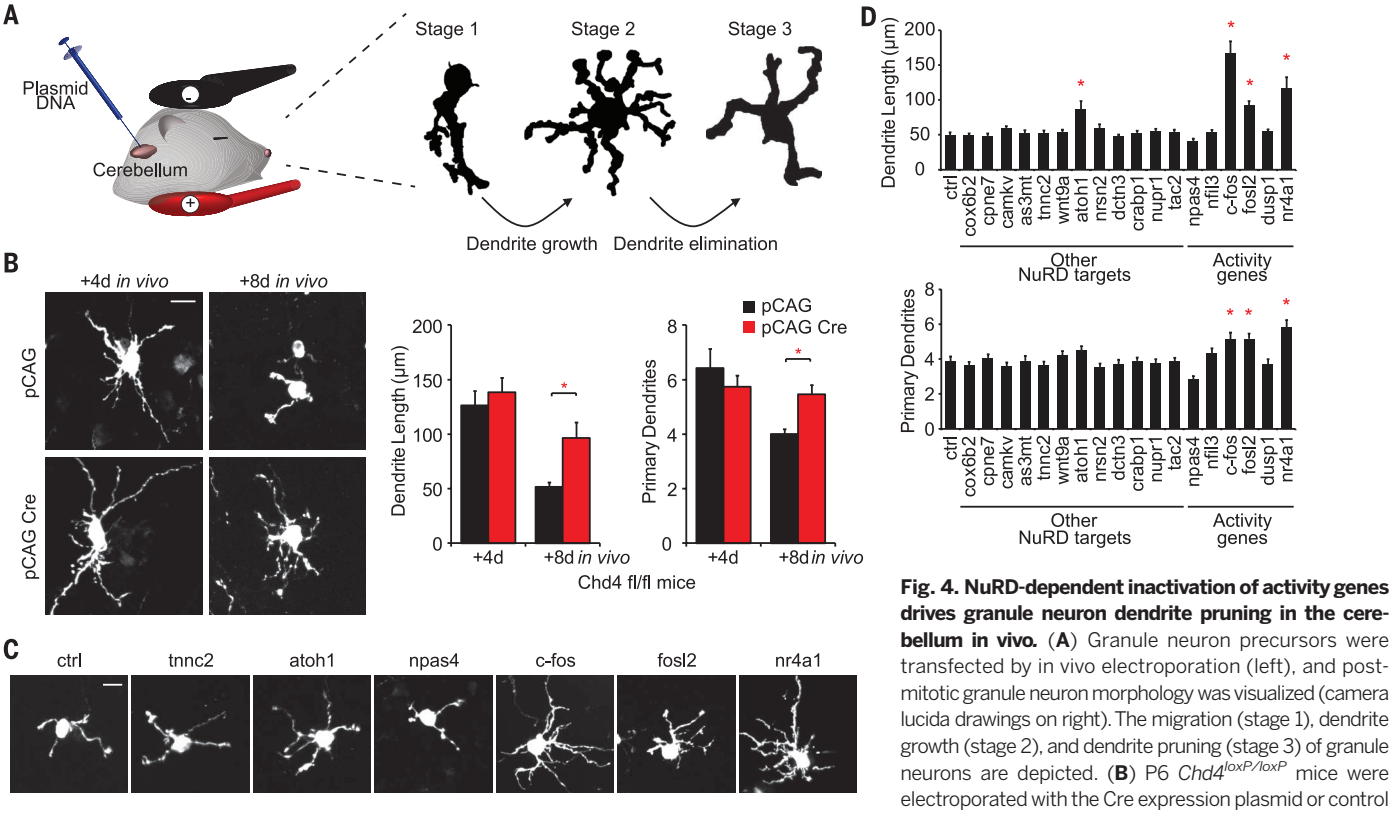



**Fig. 4. NuRD-dependent inactivation of activity genes drives granule neuron dendrite pruning in the cerebellum in vivo.** (**A**) Granule neuron precursors were transfected by in vivo electroporation (left), and postmitotic granule neuron morphology was visualized (camera lucida drawings on right). The migration (stage 1), dendrite growth (stage 2), and dendrite pruning (stage 3) of granule neurons are depicted. (**B**) P6 *Chd4*$^{loxP/loxP}$ mice were electroporated with the Cre expression plasmid or control vector together with the GFP expression plasmid and subjected to immunohistochemistry with antibody to GFP. Representative images (left) and quantification (right, *$P < 0.01$, *t* test, $n = 3$ mice). (**C** and **D**) P7 rat pups were electroporated with a panel of six activity-dependent genes or 12 other NuRD targets and analyzed as in (B). Representative images (C) and quantification [(D), *$P < 0.05$, ANOVA followed by Fisher's PLSD post hoc test, $n = 3$ to 4 rats). Scale bars: 10 μm.

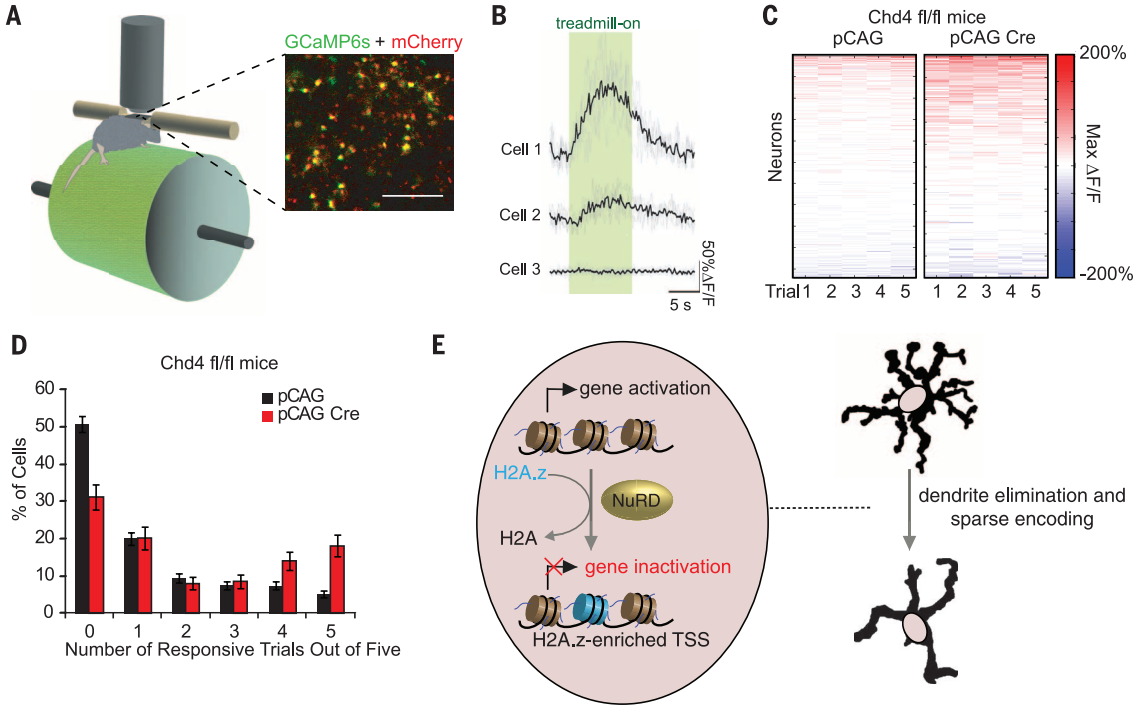



**Fig. 5. The NuRD complex promotes sparse sensorimotor encoding in behaving mice.** (**A**) Schematic of the paradigm for in vivo two-photon calcium imaging during treadmill walking. Granule neurons of lobule VI of the cerebellum electroporated with the GCaMP6s and mCherry expression plasmids (inset). Scale bar: 50 μm. (**B**) $Ca^{2+}$ transients of three representative cells in response to 10-s treadmill-on stimulus (gray: individual trials; black: trial mean; green panel: treadmill-on period). (**C** and **D**) P10–P11 *Chd4*$^{loxP/loxP}$ mice were electroporated with the Cre expression plasmid or control vector together with the GCaMP6s and mCherry expression plasmids. Heatmap of maximum change in fluorescence $\Delta F/F$ (C) during treadmill-on stimulus compared to preceding treadmill-off period for five trials per neuron in conditional Chd4 knockout and control mice and corresponding neuronal responsivity distributions (D) (Chi-Square test of independence $P < 0.00001$, $n = 5$ mice, 199 to 520 somas). Error bars represent SE from bootstrap test. (**E**) Model of the NuRD complex and H2A.z chromatin-remodeling link in the regulation of activity-dependent transcription and neural circuit assembly and function.

complex and H2A.z as epigenetic regulators that mediate the inactivation of activity genes in the brain. Thus, epigenetic mechanisms may have an active role in the inactivation of gene expression in the brain following neuronal activity.

## REFERENCES AND NOTES


1. A. E. West, M. E. Greenberg, *Cold Spring Harb. Perspect. Biol.* **3**, a005744 (2011).
2. I. Spiegel *et al.*, *Cell* **157**, 1216–1229 (2014).
3. R. N. Saha *et al.*, *Nat. Neurosci.* **14**, 848–856 (2011).
4. M. E. Greenberg, E. B. Ziff, L. A. Greene, *Science* **234**, 80–83 (1986).
5. Z. Qiu, A. Ghosh, *Neuron* **60**, 449–455 (2008).
6. D. H. Ebert *et al.*, *Nature* **499**, 341–345 (2013).
7. R. Madabhushi *et al.*, *Cell* **161**, 1592–1605 (2015).
8. L. Ho, G. R. Crabtree, *Nature* **463**, 474–484 (2010).
9. G. J. Narlikar, R. Sundaramoorthy, T. Owen-Hughes, *Cell* **154**, 490–503 (2013).
10. K. Kaji *et al.*, *Nat. Cell Biol.* **8**, 285–292 (2006).
11. Y. Rais *et al.*, *Nature* **502**, 65–70 (2013).
12. W. A. Whyte *et al.*, *Nature* **482**, 221–225 (2012).
13. T. Yamada *et al.*, *Neuron* **83**, 122–134 (2014).
14. T. Yoshida *et al.*, *Genes Dev.* **22**, 1174–1189 (2008).
15. N. Fujita *et al.*, *Cell* **119**, 75–86 (2004).
16. B. E. Bernstein *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 8695–8700 (2002).
17. H. Santos-Rosa *et al.*, *Nature* **419**, 407–411 (2002).
18. A. Barski *et al.*, *Cell* **129**, 823–837 (2007).
19. M. D. Shahbazian, M. Grunstein, *Annu. Rev. Biochem.* **76**, 75–100 (2007).
20. M. P. Creyghton *et al.*, *Cell* **135**, 649–661 (2008).
21. I. Maze, K. M. Noh, A. A. Soshnev, C. D. Allis, *Nat. Rev. Genet.* **15**, 259–271 (2014).
22. C. M. Weber, S. Henikoff, *Genes Dev.* **28**, 672–682 (2014).
23. H. S. Rhee, A. R. Bataille, L. Zhang, B. F. Pugh, *Cell* **159**, 1377–1388 (2014).
24. S. A. Denslow, P. A. Wade, *Oncogene* **26**, 5433–5438 (2007).
25. M. Heiman, R. Kulicke, R. J. Fenster, P. Greengard, N. Heintz, *Nat. Protoc.* **9**, 1282–1291 (2014).
26. G. Billings, E. Piasini, A. Lőrincz, Z. Nusser, R. A. Silver, *Neuron* **83**, 960–974 (2014).


## ACKNOWLEDGMENTS

We thank members of the Bonni laboratory for helpful discussions and critical reading of the manuscript and E. Han for advice in imaging analyses. Supported by NIH grant NS041021 (A.B.), the Mathers Foundation (A.B.), NIH grant MH093727 (J.F.M.), and NIH Medical Scientist Training Program grant T32 GM07200 (K.K.H.). We thank the Genome Technology Access Center at Washington University, which is supported by National Cancer Institute grant P30 CA91842 to the Siteman Cancer Center and by Institute of Clinical and Translational Sciences–Clinical and Translational Science Awards UL1TR000448. The authors declare no conflicts of interest. Raw sequencing data related to this study have been archived in the Gene Expression Omnibus (GEO) database under accession number GSE83253 (available at www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE83253).

## SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/353/6296/300/suppl/DC1
Materials and Methods
Figs. S1 to S10
Tables S1 and S2
References (*27*–*43*)
Movie S1

11 September 2015; accepted 12 May 2016
10.1126/science.aad4225

# Too young to win a prize in Stockholm this December?

# We don't think so.

**Science & SciLifeLab**
**PRIZE**
**FOR YOUNG SCIENTISTS**

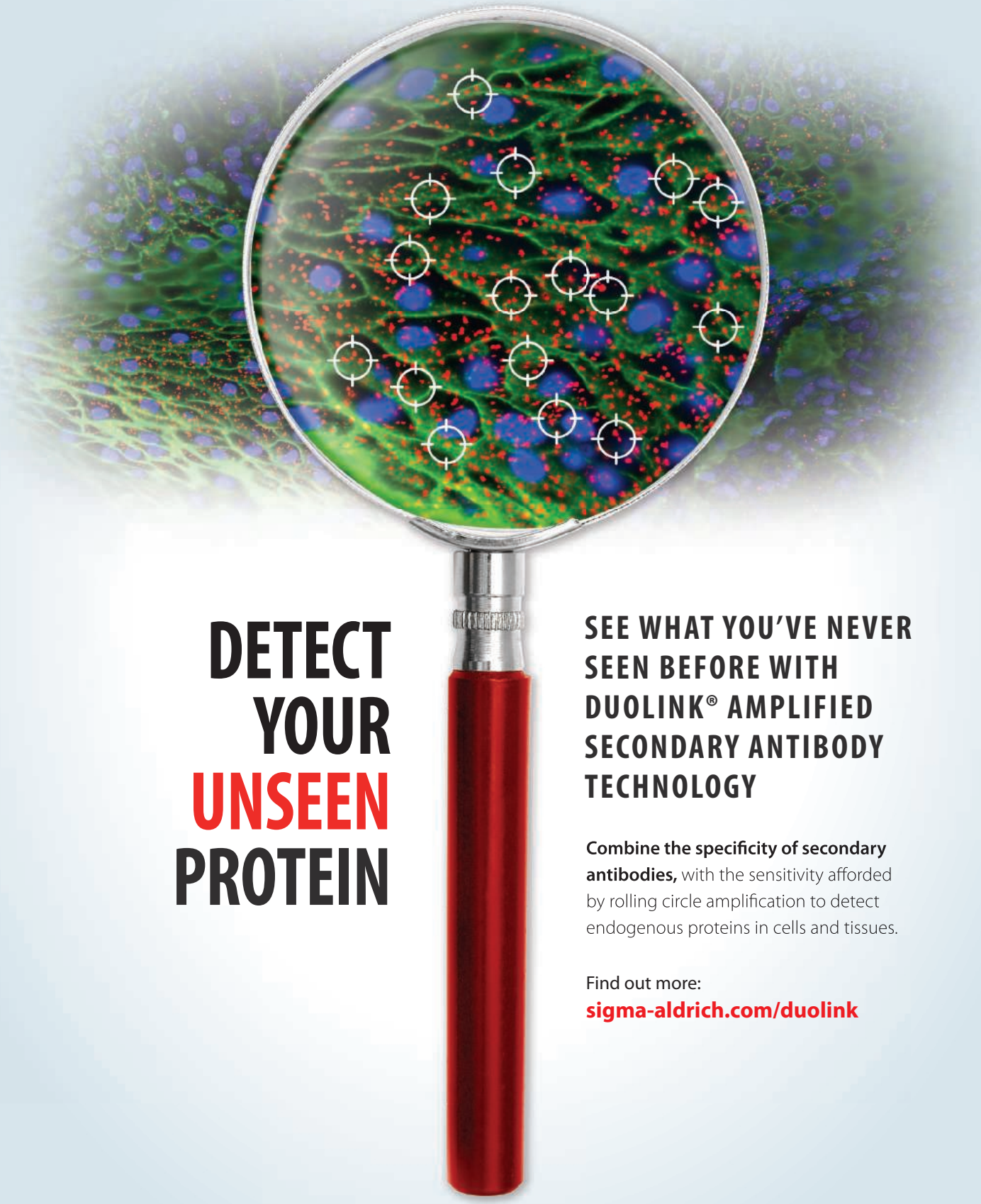
Get published in *Science* and win $30,000 USD. Enter today: www.sciencemag.org/scilifelabprize

Cell and Molecular Biology | Ecology and Environment | Genomics and Proteomics | Translational Medicine

Science AAAS

Knut och Alice Wallenbergs Stiftelse

SciLifeLab



# DETECT YOUR UNSEEN PROTEIN

## SEE WHAT YOU'VE NEVER SEEN BEFORE WITH DUOLINK® AMPLIFIED SECONDARY ANTIBODY TECHNOLOGY

**Combine the specificity of secondary antibodies,** with the sensitivity afforded by rolling circle amplification to detect endogenous proteins in cells and tissues.

Find out more:
**sigma-aldrich.com/duolink**



Duolink image credit: Caroline Welsch, Gustave Roussy Comprehensive Cancer Centre

sh9871

SIGMA-ALDRICH®

# AAAS 2017 ANNUAL MEETING

FEBRUARY 16-20 • BOSTON

SERVING SOCIETY THROUGH

# SCIENCE POLICY

**aaas.org/meetings**

Dear Colleagues:

On behalf of the AAAS Board of Directors, it is my honor to invite you to join us in Boston for the 2017 AAAS Annual Meeting, February 16-20.

The AAAS Annual Meeting is the most widely recognized global science gathering with cutting-edge scientific sessions, renowned speakers, and valuable networking opportunities.

The 2017 meeting theme—Serving Society Through Science Policy—focuses on how to inform policies with the best available scientific evidence.

We look forward to seeing you in Boston. Registration and housing open in August.

Barbara A. Schaal

**Barbara Schaal**
AAAS President
Dean of the Faculty of Arts and Sciences
Mary-Dell Chilton Distinguished Professor of Biology
Washington University in St. Louis

# Don't miss the debut of *Science Immunology*.

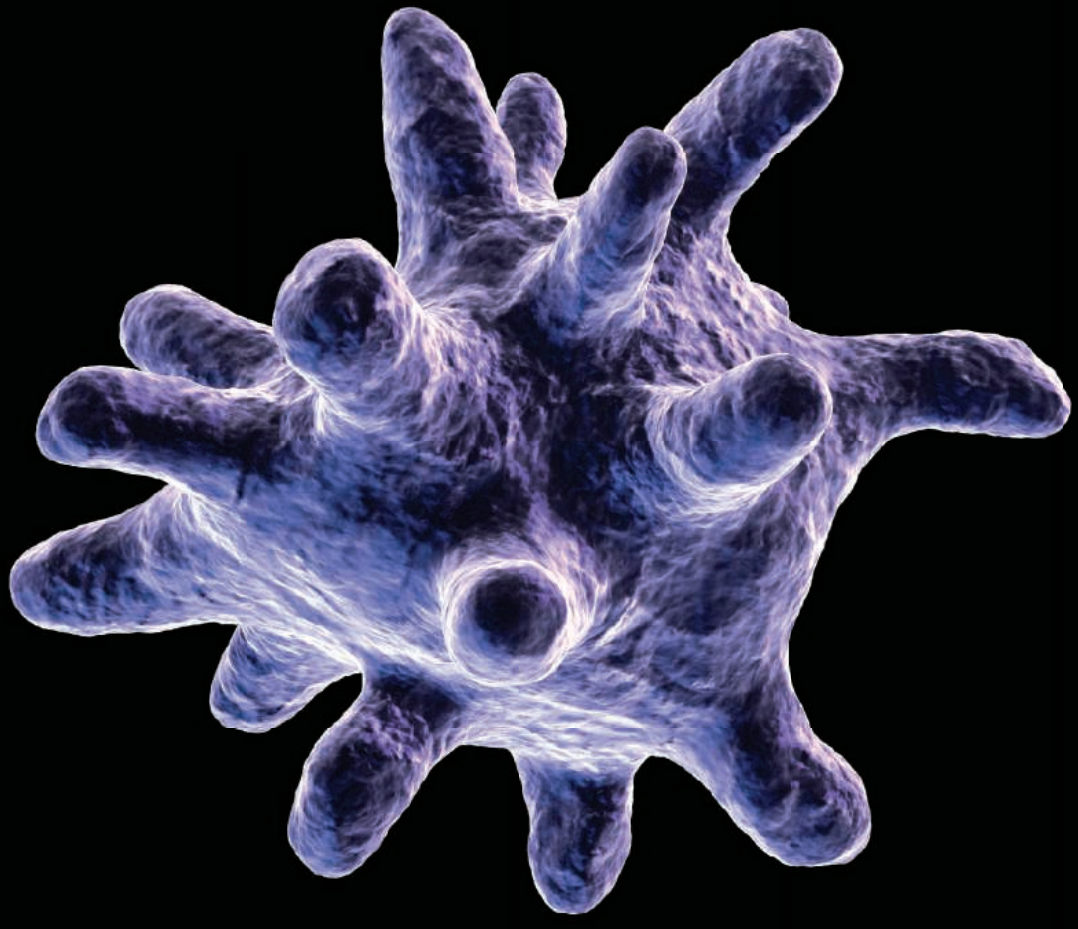


Image: Eraxion / iStockPhoto

## NOW ACCEPTING PAPERS

*Science* is expanding its reach into immunology—now offering the newest online-only, weekly journal in the *Science* family of publications. *Science Immunology* will provide original, peer-reviewed research articles that report critical advances in all areas of immunological research, including studies that provide insight into the human immune response in health and disease.

Be a part of the *Science Immunology* debut issue publishing Summer 2016!

Submit your manuscript today at **ScienceImmunology.org.**

Science Immunology

AAAS

# Structural biology shapes up

Despite advances in the field of proteomics, protein folding still remains a mystery. Yet innovations in X-ray crystallography, electron microscopy, and data analysis (think robots and Google) are yielding answers about protein structures faster than ever before. **By Alan Dove**

Within a single generation, researchers discovered that DNA was the genetic material encoding RNA, which in turn encodes proteins, and these proteins carry out the principal enzymatic reactions of life. They established that the sequences of proteins ultimately determine their 3D structures, which determine their functions. All that was left was to find a straightforward way to sequence genomes, predict the protein structures they encode, and build accurate models of all of life.

Half a century later, a few problems persist. One of the toughest is the challenge of getting from an amino acid sequence to a 3D protein structure. Though the former ostensibly determines the latter, the exact rules of the peptide-folding process remain largely inscrutable. Nonetheless, computational biologists continue to chip away at the problem, producing an evolving series of tools that can predict some classes of protein structures quite accurately. In the meantime, the classic technique of X-ray crystallography has become more accessible to nonspecialists, and a revolutionary series of developments in electron microscopy is revealing structures that previously eluded understanding.

## Advancing crystallography

Rather than trying to predict a molecule's shape from first principles, X-ray crystallographers attack the problem from the opposite direction, purifying and crystallizing a protein and then measuring how it diffracts a beam of X-rays. The technique can be tedious, often requiring thousands of experiments to determine the conditions that will yield usable protein crystals. Today, though, biologists can let robots do much of the work.

"Robots enable one to set up much smaller crystallization drops, so that you need smaller amounts of sample to screen large numbers of conditions," says Cynthia Wolberger, a professor of biophysical chemistry at **Johns Hopkins School of Medicine** in Baltimore, Maryland. Manufacturers now offer robotic systems optimized specifically for testing crystallization conditions, such as the **TTP Labtech** mosquito and the **Formulatrix** Rock Imager. Meanwhile, rapid cloning and protein expression platforms let researchers produce multiple variants of a protein to find one that will crystallize well.

Crystallographers have also improved their ability to study membrane proteins, which have been notoriously hard to crystallize. By mixing lipids, water, and proteins in specific proportions, researchers can form a cubic liquid crystal amenable to X-ray diffraction. The method requires only nanogram quantities of protein. "This is a huge advance," says Martin Caffrey, a professor of biochemistry and immunology at **Trinity College** in Dublin, Ireland. Caffrey adds that studies of G-protein coupled receptors and other critical membrane proteins "really underwent an explosion as a result of this cubic phase methodology."

Once a protein crystallizes, investigators take it to one of a few government-funded synchrotron facilities to have it bombarded with X-rays and to collect their data. That process has also gotten easier in recent years. "We rarely go [to the synchrotron] anymore; we ship our crystals there," says Wolberger, adding that "it's all been set up with great software and robotics so you can operate it from anywhere."

Improvements in synchrotron data collection systems now enable analyses that would have been impossible just a few years ago. A process called "raster scanning" allows researchers to examine crystals—and small portions of crystals—to find areas with the best diffraction in a sample that would otherwise be unusable. Automated, high-speed data collection also permits scanning many more crystals. For example, Wolberger and her colleagues recently analyzed over a thousand crystals to gather enough information for a structure, an approach that would have been impractical with manual exposures.



*Upcoming Features*

*Genomics—October 7* ■ *Neurotechniques—November 4* ■ *Cell Analysis—November 25*

### "Hit and run"

The X-ray beams themselves have improved too, with newer systems yielding higher levels of X-ray flux that enable faster, higher-resolution analyses. Meanwhile, a new technology called the "free-electron laser" has captured the attention of many structural biologists. "These are producing pulses of X-rays that are on the order of 10 to 40 femtoseconds long, so these are extremely short pulses, but there's a huge number of X-rays in each one," Caffrey explains.

That enables an entirely new type of crystallography, which Caffrey refers to as "hit and run." Researchers hit a tiny crystal with the laser's powerful X-ray pulse and collect the resulting data before the crystal explodes. While synchrotrons could already examine crystals as small as a micrometer, the new technique could enable analyses on crystals measured in nanometers.

As full-time crystallographers continue to push the limits of the technique, it's also become more accessible to researchers in other fields. That's especially true for labs working on relatively simple, soluble molecules, a description that encompasses the majority of the human genome's protein products. "I teach courses where we explain how to crystallize proteins, [and if] you attend one of these you should be able to go home and set up crystallization trials," says Caffrey.

Several vendors, including **Hampton Research, MiTeGen,** and **Molecular Dimensions** also sell ready-made protein crystallization kits, further enabling beginners to try their hand at the field. Those who succeed will find plenty of help with the next step. "There are people at synchrotrons who would be delighted to take on projects and to work with somebody at a noncrystallographic lab, and to solve a structure and become a coauthor," says Caffrey.

### Cold hard data

While X-ray crystallography has undergone a steady evolution, the other major protein structure determination technique, cryo-electron microscopy (cryo-EM), is in the midst of a revolution. In cryo-EM, researchers freeze the specimen of interest into a thin block of ice, then place it into an electron microscope, photograph it, and analyze the image to determine the target's structure. Though it's long been a useful strategy for studying relatively large structures such as viruses, technical problems limited its utility.

That suddenly changed a few years ago, thanks partly to advances that had nothing to do with the microscope itself. "By far the most dramatic [development] is new detector technology," says Eva Nogales, a professor of biochemistry and molecular biology at the **University of California, Berkeley.** Electron microscopists traditionally relied on film to capture their data, but that method came with a host of limitations. Film canisters only held 50 frames, setting a maximum number of exposures in a single sampling session, and processing, scanning, and analyzing the analog images could introduce errors.

Engineers eventually designed sensitive, radiation-hardened digital sensors that could replace film even in high-energy electron microscopes. The new sensors detect electrons directly, providing extremely high resolution with none of the inconveniences of film. "The contrast was really excellent, so we had much higher signal in those images, and they came with an extra bonus: They have very, very fast readout," says Nogales.

That development helped address another longstanding problem. The electron beam causes the frozen sample block to warp, motion-blurring a traditional camera's images. Cryo-EM researchers previously dealt with that by scanning through numerous images to find the few clear ones taken when the protein wasn't moving. With the new sensors, "the readouts are fast enough that you can take a movie during that exposure and then realign the sample to correct for that beam-induced motion, and that's had a huge impact," says David Agard, a professor of biochemistry and biophysics at the **University of California, San Francisco.**

"It's a fabulous time in the field, and there's tremendous excitement."
— David Agard

### Two angstroms and beyond

The latest generation of sensors, developed with funding from the U.S. government's economic stimulus program after the 2008 recession, can detect individual electrons. That enables cryo-EM to yield high-resolution protein structures with molecular weights of just a few hundred kilodaltons, previously the exclusive domain of X-ray crystallographers. Electron microscopists are cautiously optimistic about their ability to push further into crystallography's turf. "Whether we are going to be able to get to resolutions that are sometimes achievable by X-ray crystallography, like two [angstroms] and beyond, I do not know," says Nogales, adding that "right now the record [for cryo-EM] is around 2.2 angstroms for one very well-behaved sample."

Indeed, even dedicated cryo-EM labs still see a need for crystallography. "If you can get crystals, for the most part you get a structure immediately with crystallography, so there's no reason not to try to do that; in my lab we do that as a matter of course," says Agard. However, he adds that "there's going to be a whole realm of things that haven't crystallized well that will be taken over by cryo-EM."

Software that runs cryo-EM systems and processes their data has also improved dramatically, making everything from microscope alignment to data mining faster and easier. While the software is mostly open source, the hardware certainly isn't. Modern cryo-EM microscopes come from a single cont.>

## FEATURED PARTICIPANTS

**Centre for Integrative Systems Biology and Bioinformatics**
www.imperial.ac.uk/integrative-systems-biology

**Computation Institute, University of Chicago**
www.ci.uchicago.edu

**Direct Electron**
www.directelectron.com

**FEI**
www.fei.com

**Formulatrix**
www.formulatrix.com

**Gatan**
www.gatan.com

**Hampton Research**
www.hamptonresearch.com

**Johns Hopkins School of Medicine**
www.hopkinsmedicine.org

**MiTeGen**
www.mitegen.com

**Molecular Dimensions**
www.moleculardimensions.com

**Trinity College Dublin**
www.tcd.ie

**TTP Labtech**
www.ttplabtech.com

**University of California, Berkeley**
www.berkeley.edu

**University of California, San Francisco**
www.ucsf.edu

**University of Chicago**
www.uchicago.edu

## ADDITIONAL RESOURCES

**CAMEO**
www.cameo3d.org

**Phyre2**
www.sbg.bio.ic.ac.uk/phyre2

**RaptorX**
raptorx.uchicago.edu

company, **FEI,** and sensors for them are made by it, **Direct Electron,** and **Gatan.** Regardless of which vendors and instruments one chooses, a complete cryo-EM system costs several million dollars to buy, and several hundred thousand annually to maintain and operate. "Only the richest labs or institutions that are subsidizing this will be able to do this, and that's not a good situation in terms of national capabilities," says Agard. Germany, the United Kingdom, and China have all established national-level funding for cryo-EM facilities, while the U.S. National Institutes of Health is still deciding how to proceed.

### Siri, what does the protein look like?

Regardless of how fast X-ray crystallography and cryo-EM develop, though, neither is likely to satisfy researchers' growing desires for structural models. "You can't beat an experimentally determined structure, [but] the challenge is the vast number of genomes that have been sequenced, and the effectively impossible task of devising 3D structures for all the proteins," says Michael Sternberg, director of the **Centre for Integrative Systems Biology and Bioinformatics** at Imperial College London, United Kingdom.

To address that challenge, Sternberg and his colleague Lawrence Kelley have focused on the decades-old dream of molecular biologists: predicting protein structures directly from sequence information. The team's latest tool, Phyre2, uses template-based modeling, a process that compares a sequence against a global database of previously solved protein structures. Several other research groups have developed similar tools, which are accessible online to researchers worldwide.

These tools can yield impressive results, at least for protein families that are well-represented in the databases. "On some protein samples we can do pretty well, since now we have much more data, so we can apply or develop much more sophisticated techniques to model the sequence–structure relationship," says Jinbo Xu, a senior fellow of the **Computation Institute** at the University of Chicago. Xu is the principal investigator behind RaptorX, another online portal for predicting secondary and tertiary protein structures.

In addition to template-based modeling, some researchers are now exploring a technique called "contact prediction." This approach searches through vast troves of sequence data to identify evolutionarily conserved amino acid interactions, then uses those correlations to predict a novel sequence's folding patterns. "That's certainly proving very useful for membrane-bound proteins where there are very few crystal structures available," says Sternberg, but he adds that "you still need quite a large number of aligned sequences" for contact prediction to work.

Besides improving the underlying algorithms, computational biologists have been working on making their system interfaces friendlier. "We're very much following the sort of Google approach of a very clean screen, not inundating the user with many options, and really only delivering what the user wants," says Sternberg. It seems to be working. Sternberg estimates that of the 44,000 unique users who accessed Phyre2 last year, only a few hundred contacted him or Kelley for support.

### Expanding options

Xu and his colleagues have also embraced the user-friendly model, with similar results. "The broader community is using the tools, [and] the users of my server have very diverse backgrounds," he says. Researchers without structural biology training may not understand the limitations of the underlying algorithms, so most of the major portals are designed to help users interpret their results. RaptorX provides a quality evaluation along with each protein model, scoring how likely the structure is to be correct. Similarly, Phyre2 provides an overall confidence score, as well as individual scores for substructures down to the amino acid level.

Because all of the major structure prediction tools are free online, scientists can also hedge their bets by sending their target sequences to all of them to see how the models differ. Another site, **CAMEO,** also tests and rates the different protein-structure services. Each week, CAMEO sends a test sequence to all of the participating servers, then calculates benchmarks based on the systems' speed, server reliability, and other characteristics.

Different tools also have distinct strengths and weaknesses depending on a researcher's specific needs. RaptorX emphasizes template-based modeling for particularly challenging proteins, while other portals are better suited for rapid analysis of large numbers of relatively well-characterized protein families. Phyre2 offers a suite of additional services; the "Phyre Alarm," for example, will send an email to researchers when new data allow the system to calculate an improved model of their protein.

Regardless of the approach they choose, investigators seeking protein structures should expect more good news. Surveying the progress in all three techniques, Agard echoes the general sentiment of structural biologists: "It's a fabulous time in the field, and there's tremendous excitement."

*Alan Dove is a science writer and editor based in Massachusetts.*

DOI: 10.1126/science.opms.p1600107

### NMR Magnet and Probe

A nuclear magnetic resonance (NMR) system has been equipped with an actively shielded Aeon 1-GHz magnet at the University of Bayreuth, Germany, while a single-story Aeon 950-MHz magnet has been successfully installed at the University of Leeds, United Kingdom. Active shielding reduces the space requirements for the two-story magnet by more than an order of magnitude. The Aeon magnets feature novel active refrigeration technology, eliminating the need for liquid nitrogen and reducing liquid helium boil-off essentially to zero under normal operation. In conjunction with the latest GHz magnets and novel nitrogen-15 ($^{15}$N) direct-detect NMR methods, the CryoProbe now makes direct $^{15}$N detection sensitive and advantageous in very large globular proteins and intrinsically disordered proteins, due to the longer relaxation times, high resolution, and low chemical shift anisotropy of $^{15}$N spectra. $^{15}$N detection is beneficial in cases where carbon-detected methods suffer from multiple couplings to neighboring carbons, or in the study of proline-rich protein domains and paramagnetic metalloproteins.

**Bruker**
For info: +49-(0)-721-5161–6500
www.bruker.com

### UHPLC Columns

Luna Omega 1.6-µm particle UHPLC columns deliver improved performance compared to other sub-2-µm products on the market, along with exceptionally tight particle size distribution, high column packing efficiencies, steady selectivity, and excellent reproducibility. Luna Omega columns are well suited for drug discovery and development, food contaminant analysis, environmental testing, toxicology, and clinical research. The columns feature novel silica with a constant particle morphology that is modified with proprietary, postsynthetic thermal treatment processes for mechanical strength and inertness. These beads provide significantly better peak shape for compounds of interest and minimize unwanted secondary interactions. Proprietary bonding technologies ensure wide stationary phase coverage and excellent separation power. Together, Kinetex core-shell technology and Luna Omega columns provide an ideal, complementary UHPLC solution. Combinations such as the Luna Omega 1.6-µm C18 with the Kinetex 1.7-µm Biphenyl and Kinetex 1.7-µm F5 promote greater method development success.

**Phenomenex**
For info: 310-212-0555
www.phenomenex.com


### Gas Purifiers

Gas purity is essential in any GC, HPLC, or GC/MS application requiring high sensitivity. Available in optimized models for helium, hydrogen, nitrogen, and pure air, GasTrap gas purifiers dramatically reduce the levels of contaminants, enhance the purity of lower grade helium, help ensure instrument stability and reproducibility, and lower running costs. Laboratory gases for GC, HPLC, and GC/MS are typically supplied in pressurized cylinders that contain trace impurities. Even with specialty high-purity gas supplies, trace contaminants of up to 20 ppm oxygen, water, carbon dioxide, neon, and argon are present, which increase proportionately as the cylinder pressure drops. Unlike disposable gas filters that are time consuming, costly, and cannot remove gases such as neon, argon, or nitrogen, GasTrap gas purifiers use patented Pressure Swing Adsorption technology to remove all contaminant trace gases and self-clean every cycle.

**Asynt**
For info: +44-(0)-1638-781709
www.asynt.com

### Running Buffer

The new Fisher BioReagents Fast*Run* Tris SDS PAGE Running Buffer provides comparable or better resolution and an increased molecular weight separation range of proteins when compared to traditional Tris-glycine-sodium dodecyl sulfate (SDS) buffer. The new buffer system also reduces the number of Tris-glycine polyacrylamide gel compositions required to resolve a protein, saving researchers time and money. Tris-glycine mini-gels (precast or homemade) prepared with the conventional buffer system are typically run at 125 V. The buffer system heats up at high voltages, which in turn heats the gel, resulting in a breakdown of the protein bands and loss of resolution. The Fisher BioReagents Fast*Run* buffer can be run at a higher voltage (200 V recommended), because it does not generate excessive heat. The result is a significant improvement in run times. In an example, gels run with the new buffer system took only 25 minutes, compared to 90-minute run times with the traditional reagents.

**Fisher Scientific**
For info: 800-766-7000
www.fishersci.com

### Therapeutic Protein Analysis Kits, Columns, and Software

Biopharmaceutical companies can now achieve advanced levels of specificity, sensitivity, and speed with these specialized analytical workflow solutions for the characterization and quantification of proteins. Thermo Scientific DNAPac RP columns separate and characterize DNA/RNA oligonucleotides and longer double-stranded DNA fragments (up to 10k base pairs) using reversed-phase (RP) HPLC and LC/MS. The Thermo Scientific SMART Digest Kit is a protein digestion kit for reproducible, sensitive, and fast bottom-up analysis of proteins. The Thermo Scientific Vanquish Flex UHPLC System is a fully bioinert UHPLC offering that delivers reliable separations. The Thermo Scientific Dionex Chromeleon 7.2 Chromatography Data System answers questions that frequently arise during method development. The Thermo Scientific MAbPac RP column conducts high-resolution separation of intact proteins, such as monoclonal antibody (mAb) variants and antibody drug conjugates. Thermo Scientific BioPharma Finder software provides intelligent batch/lot comparisons to promote consistency and quality throughout.

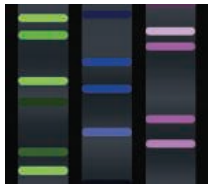
**Thermo Fisher Scientific**
For info: 800-955-6288
www.thermofisher.com

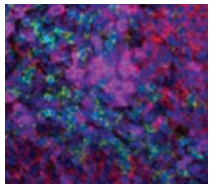
Electronically submit your new product description or product literature information! Go to www.sciencemag.org/about/new-products-section for more information.

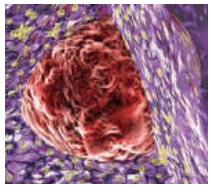
Newly offered instrumentation, apparatus, and laboratory materials of interest to researchers in all disciplines in academic, industrial, and governmental organizations are featured in this space. Emphasis is given to purpose, chief characteristics, and availability of products and materials. Endorsement by *Science* or AAAS of any products or materials mentioned is not implied. Additional information may be obtained from the manufacturer or supplier.

# want new technologies?

antibodies
apoptosis
biomarkers
cancer
cytometry
data
diseases
DNA
epigenetics
genomics
immunotherapies
medicine
microbiomics
microfluidics
microscopy
neuroscience
proteomics
sequencing
toxicology
transcriptomics







## watch our webinars

Learn about the latest breakthroughs, new technologies, and ground-breaking research in a variety of fields. Our expert speakers explain their quality research to you and answer questions submitted by live viewers.

**VIEW NOW!**
**webinar.sciencemag.org**

Science
AAAS

Brought to you by the *Science*/AAAS Custom Publishing Office

@SciMagWebinars


There's only one Science

# *Science* Careers Advertising

For full advertising details, go to ScienceCareers.org and click For Employers, or call one of our representatives.

**Tracy Holmes**
Worldwide Associate Director
*Science* Careers
Phone: +44 (0) 1223 326525

THE AMERICAS

E-mail: advertise@sciencecareers.org
Fax: +1 (202) 289 6742

**Tina Burks**
Phone: +1 (202) 326 6577

**Nancy Toema**
Phone: +1 (202) 326 6578

**Online Job Posting Questions**
Phone: +1 (202) 312 6375

EUROPE / INDIA / AUSTRALIA / NEW ZEALAND / REST OF WORLD

E-mail: ads@science-int.co.uk
Fax: +44 (0) 1223 326532

**Sarah Lelarge**
Phone: +44 (0) 1223 326527

**Kelly Grace**
Phone: +44 (0) 1223 326528

**Online Job Posting Questions**
Phone: +44 (0) 1223 326528

JAPAN

**Katsuyoshi Fukamizu** (Tokyo)
E-mail: kfukamizu@aaas.org
Phone: +81 3 3219 5777

**Hiroyuki Mashiki** (Kyoto)
E-mail: hmashiki@aaas.org
Phone: +81 75 823 1109

CHINA / KOREA / SINGAPORE / TAIWAN / THAILAND

**Ruolei Wu**
E-mail: rwu@aaas.org
Phone: +86 186 0082 9345

**Danny Zhao**
E-mail: dzhao@aaas.org
Phone: +86 131 4114 0012



ScienceCareers
FROM THE JOURNAL SCIENCE AAAS

ScienceCareers.org

## myIDP: A career plan customized for you, by you.


For your career in science, there's only one Science

my IDP ScienceCareers

**Recommended by leading professional societies and the NIH**

**Features in myIDP include:**

- Exercises to help you examine your skills, interests, and values.
- A list of 20 scientific career paths with a prediction of which ones best fit your skills and interests.
- A tool for setting strategic goals for the coming year, with optional reminders to keep you on track.
- Articles and resources to guide you through the process.
- Options to save materials online and print them for further review and discussion.
- Ability to select which portion of your IDP you wish to share with advisors, mentors, or others.
- A certificate of completion for users that finish myIDP.

Visit the website and start planning today!
myIDP.sciencecareers.org

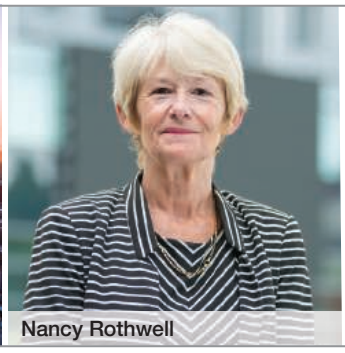
ScienceCareers AAAS In partnership with:

advancing the life sciences 1912-2012 FASEB Federation of American Societies for Experimental Biology

Medical College of Wisconsin

UCSF University of California San Francisco

Burroughs Wellcome Fund



Nancy Rothwell

Konstantin Novoselov

# Find the best of both worlds in Manchester

At the 2016 EuroScience Open Forum (ESOF), Manchester will flaunt its stuff. With over £500 million (US$720 million) flowing in from both public and private sources to support new science facilities and research, the city has lots to show off. For job seekers, Manchester offers an atmosphere where industry works seamlessly with academia, making it possible to have a foot in both worlds. **By Gunjan Sinha**

**Melissa Denecke** was anxious. She had spent the past 30-plus years in Germany—her entire postgraduate and academic career—successfully conducting nuclear-related research. Then in 2011, the Fukushima catastrophe happened. The 15-meter tsunami that hit Japan damaged several nuclear power plants, which then leaked radioactive contamination into the environment. Shortly afterwards, the future of nuclear research in Germany turned grim. The government announced a phase-out of all nuclear power by 2022. "There was a department meeting where I felt very apprehensive about future funding prospects," she recalls. At the time, Denecke was at the Karlsruhe Institute of Technology in Germany, where she was quite content as a department head. But the political fallout from Fukushima made her consider a new home. In 2012, The University of Manchester offered her a job as codirector of its Dalton Nuclear Institute. She jumped at the chance.

At a time when many institutions are scaling back nuclear research, The University of Manchester is forging ahead. The effort is part of a broader initiative to grow the regional economy by identifying the city's strengths and capitalizing on them, says **Michael Contaldo** of New Economy Manchester—an organization created to help Greater Manchester's economy grow. "Science is a key area," he says. Two years ago the university identified five broad research areas in which it already excelled and classified them as "research beacons." Those beacons are energy (including nuclear), advanced materials, industrial biotechnology, cancer, and addressing global inequalities. The beacons highlight areas of excellence and impact within a diverse range of inquiry to give focus and demonstrate "how some of our best research is being used to solve global challenges," says Professor Dame **Nancy Rothwell,** the university's president and vice-chancellor. "We asked ourselves, 'What is Manchester excellent at?' We needed to define that."

In all five beacon areas, major developments are either already in the works or planned for the near future—developments that will strengthen the authority and reputation of both the university and Manchester itself. The initiative places a strong emphasis on application and translation: Industry is either already involved or will be heavily courted. Manchester already has a strong history of collaboration with industry, says Rothwell. By focusing on the research beacons, the university hopes to attract even more investment to the city. Outsiders have taken note of its massive investment in science: ESOF, Europe's largest scientific meeting, chose Manchester as the European City of Science for 2016, and the city will host the biennial ESOF meeting this summer.

For job seekers, synchronicity between academia and industry means that students and researchers have many opportunities to experience both worlds, Rothwell adds. Manchester's expansion as a technological science hub also brings more opportunities than ever to study or work in the city.

## Energy research

Manchester's prominence in nuclear research spans over two centuries: John Dalton, Ernest Rutherford, and Niels Bohr—to name just a few—all conducted significant research there that helped form modern theories of chemistry and physics. October 2016 will mark the 60th anniversary of the opening of the world's first civil nuclear energy power plant, which was located in northwest England at Calder Hall and decommissioned in 2003. That legacy of nuclear energy development continues today at The University of Manchester in the form of research on new reactor technologies and related areas, says Denecke.

At the Dalton Nuclear Institute, Denecke "wears two hats," she says. She is now the scientific director of Dalton, where she is responsible for coordinating nuclear research at the university; she also holds a chair in chemistry. The university hosts a broad range of nuclear-related research that not only encompasses the operation and decommissioning of nuclear power plants and their fuels, but also cuts across various other disciplines including health, medicine, and the environment. For example, researchers in biogeochemistry study how microbes interact with radioactive elements in the environment—especially long-lived elements such as uranium, plutonium, and neptunium. Other researchers build robots that can be sent to investigate nuclear disaster sites where the status of the radioactive components is unknown.

A huge advantage of carrying out nuclear-related research at Manchester is the university's proximity to Calder cont.>

PHOTOS: (LEFT) ©VISITENGLAND/MARKETING MANCHESTER; (RIGHT TWO) MATT THOMAS/THE UNIVERSITY OF MANCHESTER

## *Upcoming Features*

*Postdoc Careers—August 26* ■ *Faculty Careers—September 16* ■ *Faculty Careers—October 7*


Nigel Scrutton


Melissa Denecke

**"People come in from different disciplines and they need time to form a community that grows."**
**—Nigel Scrutton**

Hall and Sellafield—a nuclear fuel reprocessing and decommissioning site, and home to the United Kingdom's National Nuclear Laboratory Central Laboratory. These sites provide opportunities to study technologies and processes in the real world. In addition, says Denecke, "We have facilities that are not available elsewhere, like the Dalton Cumbrian facility"—a research facility connected to the Dalton Nuclear Institute that hosts national nuclear research activities. Considered in combination, these resources comprise a vast and almost unparalleled array of equipment available to researchers, she adds.

Through collaboration with industrial partners such as Areva and EDF Energy, the Dalton Nuclear Institute also trains students to work in the nuclear industry. A critical mass of the nuclear workforce will retire soon, says Denecke. The University of Manchester, in collaboration with other universities, industrial partners, and various other agencies, sponsors training centers where students can earn doctoral degrees in areas that target such skill deficits as next-generation nuclear technologies or materials for demanding environments. For anyone looking for an education or a job, "there's a huge range of things that we do," says Denecke. "There are even opportunities in the humanities, such as studying the anthropology of technology."

Nuclear-related research activities are set to expand. Space in the new Sir Henry Royce Institute for Advanced Materials Research—a £235 million (US$338 million) publicly funded facility scheduled to open in 2019 that will be supported by six UK partner universities—is already dedicated to research on nuclear technologies. Among the new center's goals is to accelerate the use of advanced materials in existing and emerging industrial sectors.

## Advanced materials research

Another materials research center, the newly opened National Graphene Institute (NGI), aims to capitalize on Andre Geim and **Konstantin Novoselov's** isolation of graphene from graphite in 2004, which took place at The University of Manchester and for which they won a Nobel Prize in 2010. At NGI, academics work with industrial partners to investigate new applications of 2D, atomic-scale material. Novoselov, a professor of physics at the university, was very involved in the design of the NGI's new facility. According to Novoselov, the facility was designed with three goals in mind: to expand existing research on graphene, to capitalize on that research by commercializing applications, and to take research on atomic-scale materials beyond just graphene.

**Rahul Raveendran-Nair,** for example, is developing graphene-based membranes. A few years ago, his lab made graphene oxide—a functional form of graphene—and fabricated it into a multilayer, micrometer-thick, paper-like membrane. The membrane allows only water to pass through, but no other solvents, "not even helium atoms, which are the smallest atoms," he explains. The membrane's most obvious application is to filter water, but it could be developed for other industrial applications as well, he adds.

Raveendran-Nair left India to join a Ph.D. program at The University of Manchester in 2007. He had been studying carbon nanotubes while in India and seized the opportunity to work with graphene pioneers Geim and Novoselov. For his doctoral thesis, he studied the physical properties of graphene, which turned into a landmark paper published in *Science*. "It was a really successful project that gave me a lot of motivation," he says. Working with other researchers, he captured the first image of the atomic structure of graphene and synthesized two new chemical derivatives, graphane and fluorographene. He now runs his own lab and was recently promoted to full professor.

For anyone interested in graphene research, "in terms of equipment, we probably have the best in the world," says Novoselov. In an ironic twist on the scientific method, however, in his case equipment played no role at all—he isolated graphene using double-sided tape! Nevertheless, his no-tech method of achieving this Nobel-Prize winning result is the exception rather than the rule, he says. Equally important in promoting award-winning science is collaboration, he adds. At the NGI, Novoselov has worked diligently to use architecture to create a "collegial" environment that facilitates interaction. The effect, he hopes, will be faster results. "You don't want to have a slow start in redoing what people have done already."

The UK government contributed £38 million (US$55 million) toward the construction of NGI, as part of £50 million (US$72 million) allocated to graphene research; an additional £23 million (US$33 million) came from the European Regional Development Fund (ERDF). Another materials research facility scheduled to open at the university in 2017 is the £60 million (US$86 million) Graphene Engineering Innovation Centre. It has been financed through a combination of public and private funds. The centre's goal is to accelerate applied R&D of advanced materials in partnership with other research organizations and industry.

## Industrial biotechnology

Taking **Nigel Scrutton's** experience at the Manchester Institute of Biotechnology (MIB) as a measure, it takes several years to develop a truly interdisciplinary culture. The MIB opened in 2006, and was the first "university-based, purpose-built interdisciplinary research institute of its kind in the UK," according to the MIB website. Its aim was to come up with biotechnological solutions to problems at the interface of medicine, biology, physical sciences, engineering, mathematics, and computation. The result was researchers from disparate scientific disciplines working together on large, collaborative projects. "It took time to embed a common sense of purpose," says Scrutton, who has been at the MIB since it opened. "These things don't happen from day one. People come in from different disciplines and they need time to form a community that grows."

It was into this well-established interdisciplinary environment that **Eriko Takano** arrived in 2012 to lead synthetic biology research at the MIB. She came from the University of Groningen in the Netherlands, where she had been an associate professor. She was drawn to the MIB specifically because of the environment. "It's time-consuming to make contact with colleagues who aren't around you. Here you can talk to the right people immediately," she says. She also enjoys the collaborative work ethos at MIB as well as the welcoming attitudes toward women.

PHOTOS: MATT THOMAS/THE UNIVERSITY OF MANCHESTER

## FEATURED PARTICIPANTS

Cancer Research UK Manchester Institute
www.cruk.manchester.ac.uk

Dalton Nuclear Institute
www.dalton.manchester.ac.uk

EuroScience Open Forum
www.esof.eu

International Centre for Advanced Materials
www.icam-online.org

Manchester Cancer Research Centre
www.mcrc.manchester.ac.uk

Manchester Institute of Biotechnology
www.mib.ac.uk

New Economy Manchester
neweconomymanchester.com

The Christie
www.christie.nhs.uk

The University of Manchester
www.manchester.ac.uk

Eriko Takano

Rahul Raveendran-Nair

**"It's time-consuming to make contact with colleagues who aren't around you. Here you can talk to the right people immediately."**
**—Eriko Takano**

"You really feel like ideas that may be different from the norm are heard and supported."

Takano uses synthetic biology to create microbes capable of producing novel antibiotics and other chemicals. Her team of scientists includes experts in biochemistry, engineering, informatics, and social science.

Public perception of biotechnology in Western Europe remains negative. "We are looking at how we can be really responsible about our innovative technologies and how we can make people aware of the potential benefits of what we are doing," Takano says. "Synthetic biology could potentially revolutionize industrial biotechnology, and MIB is the perfect place to do this research because it is so interdisciplinary."

### Cancer research

Manchester has brought the translation from bench to marketplace into life science research as well, particularly in cancer research. **Dan Wiseman**'s career exemplifies how integrating basic research and clinical science benefits both researchers and patients. Wiseman is a clinician scientist specializing in acute myeloid leukemia. His lab is located in the Cancer Research UK Manchester Institute, in a building that sits adjacent and is connected to The Christie—the largest single-site cancer center in the United Kingdom, where more than 44,000 patients a year are treated.

The proximity of the two buildings enables Wiseman to treat patients and also use their samples for his research. In his lab, Wiseman studies whether the oncometabolite 2-hydroxyglutarate (2HG), which has been shown to drive leukemic cells to proliferate, can serve as a biomarker to both monitor the disease and its response to treatment. Cells secrete 2HG when the isocitrate dehydrogenase 1 (*IDH1*) or *IDH2* genes are mutated. He also investigates the effects of 2HG inhibitors on cancer cells.

"I can identify patients with these mutations in real time, obtain material for the lab, and take it to the research institute on the same day," Wiseman explains. He says obtaining his own samples allows him to gain a detailed understanding of the clinical context of each individual patient sample, "rather than be at the mercy of the information that has been provided." He is also directly involved with the patients. "It's a great model for integrating clinical and translational research."

The Manchester Cancer Research Centre (MCRC), which moved to a new purpose-built facility last summer, will intensify translational cancer research in the city even more. The new building is a £28.5 million (US$41 million) partnership between Cancer Research UK, The University of Manchester, and The Christie NHS (National Health Service) Foundation Trust, and sits across the street from The Christie. Like the other new buildings that have resulted from Manchester's research initiative, the MCRC is designed to promote collaboration. It will house 150 researchers and clinician scientists, as well as an additional 100 people working in clinical trial design and administration. The Christie has the largest phase 1 clinical trials unit in the world, and its cancer researchers also work closely with industry to take their drugs through trials.

Such an open-arms reception of industry benefits job seekers in myriad ways, says **Robert Sorrell,** British Petroleum's (BP's) vice president of public-private partnerships. Researchers working at the intersection of both worlds "have a much better appreciation of the challenges that industry faces. They can also make more informed decisions as to whether they want to work in industry," he says.

In 2012, BP launched the International Centre for Advanced Materials, with an investment of US$100 million over 10 years. The center is an initiative spearheading research collaboration between The University of Manchester and three other universities, in technologies related to the oil and gas industry. "We have really been impressed with the university's skills in project management and in forming these highly effective collaborations. They really understand what industry needs," Sorrell says. The initiative includes financial support for students and professors; and all projects require a BP mentor, providing academic collaborators with even closer contact to industry.

Manchester's breadth of scientific prowess is unfortunately far less renowned than its soccer teams and music scene, laments Rothwell.

Through the 2016 ESOF City of Science designation, however, she hopes that perceptions will start to change. Indeed, the motto for ESOF is "Science as Revolution," an overarching theme that aims to encourage debate and exploration of how science and technology transforms our lives, as well as a nod toward Manchester's heritage as a science city. The university plans to showcase its cutting-edge research and the opportunities available for anyone to do science in Manchester. "This is a great European city and a fantastic place for science," says Rothwell. "I want people to see and remember that."

*Gunjan Sinha is a freelance writer living in Berlin, Germany.*

DOI: 10.1126/science.opms.r1600165

PHOTOS: MATT THOMAS/THE UNIVERSITY OF MANCHESTER

# THE KAVLI FOUNDATION

## WE ARE SEEKING SCIENCE PROGRAM OFFICERS

**Do you have a strong desire to create a positive impact in the world through science?**

**Are you interested in harnessing the entirety of your talents -- beyond the bench -- to advance science?**

**If so, we'd like to hear from you. We're building a team of enthusiastic scientists who are excited to think big and catalyze new ideas for the future of science.**

**How do we push science forward? We work as a team to ensure the success of our global network of Kavli Institutes in Science. We work closely with scientists to understand gaps and opportunities to accelerate scientific progress. Our work has played a central role in the development of collaborative national movements in science, including The Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative and the National Microbiome Initiative.**

**Come join us. Dare to make big differences in science.**

DUTIES AND RESPONSIBILITIES

1. Strategic planning
2. Monitoring and evaluating programs
3. Gathering and analyzing data
4. Participate in and organize conferences, events and visit institutes – 25% travel
5. Special projects
6. Interact with scientists, government and business leaders

QUALIFICATIONS

- Ph.D. in Physical or Life Sciences with 7 years of experience minimum
- Experience with higher education institutions and non-profits
- Ability to synthesize and analyze information
- Scientific curiosity and willingness to learn other scientific disciplines
- Develop and facilitate relationships with Foundation's stakeholders and colleagues
- Strong oral, written and people skills
- Specialty in neuroscience, nanoscience, theoretical physics or astrophysics

BENEFITS

- Medical and Dental
- Life Insurance with AD&D
- Short Term and Long Term Disability plans
- Paid Vacation, Holidays, and Sick Leave
- 401(k) plan with a generous match

For more information, please visit: www.kavlifoundation.org/jobs. To apply, please submit your curriculum vitae to: hr@kavlifoundation.org

---

EGL Charitable Foundation

The EGL Charitable Foundation invites you to apply to the

## Gruss Lipper Post-Doctoral Fellowship Program

### Eligibility

- Israeli citizenship
- Candidates must have completed PhD and/or MD/PhD degrees in the Biomedical Sciences at an accredited Israeli University/Medical School or be in their final year of study
- Candidates must have been awarded a postdoctoral position in the U.S. host research institution.

**Details regarding the fellowship are available at www.eglcf.org**

NW210508R

---

UNIVERSITY OF CAMBRIDGE

**www.jobs.cam.ac.uk**

## The Al-Kindi Professorship

Department of Genetics • Ref: PC09542

The Board of Electors to the Al-Kindi Professorship invite applications for this Professorship from persons whose work is connected with the quantitative dynamics of living systems to take up appointment on 1 October 2017, or as soon as possible thereafter.

Candidates will have an outstanding research record of international stature in an interdisciplinary area concerning quantitative approaches to biological systems. They will have the vision, leadership, experience and enthusiasm to build on current strengths in genetics and physical biology while maintaining and developing a leading research presence and contributing to the strategic development of the discipline within Biological Sciences and more widely across the University. They will hold a PhD or equivalent postgraduate qualification.

Standard professorial duties include teaching and research, examining, supervision and administration. The Professor will be based in Cambridge.

A competitive salary will be offered.

**To apply online for this vacancy and to view further information about the role, please visit: http://www.jobs.cam.ac.uk/job/10801.**

**Closing Date: 05 September 2016**

**Further information is available at: www.admin.cam.ac.uk/offices/academic/secretary/professorships/ or contact the Academic Secretary, University Offices, The Old Schools, Cambridge, CB2 1TT, (email: ibise@admin.cam.ac.uk).**

**Applications, consisting of a letter of application, a statement of current and future research plans, a curriculum vitae and a publications list, along with details of three referees should be made online no later than Monday 5 September 2016.**

**Informal enquiries may be directed to Professor Anne C Ferguson-Smith, Head of the Department of Genetics, Cambridge, telephone +44 (0)1223 333834 or e-mail afsmith@gen.cam.ac.uk.**

**Please quote reference PC09542 on your application and in any correspondence about this vacancy.**

The University values diversity and is committed to equality of opportunity.

The University has a responsibility to ensure that all employees are eligible to live and work in the UK.

AgreenSkills plus

*Pathways for inventive researchers*

## AgreenSkills+ Postdoctoral Research Fellowships

**Next application deadlines: 14 October 2016 & 28 April 2017**

AgreenSkills+ is an international postdoc fellowship programme co-funded by the EU within the FP7 Marie Curie Actions, accepting proposals from experienced young and independent researchers in the fields of **life and environmental sciences** with up to 10 years' experience after their doctorate. Proposals are accepted, without any guideline topics, for **basic or applied sciences**.

AgreenSkills+ programme is open to incoming researchers who are offered 1 to 3 years fellowships to carry out a research project in the best labs and research organisations in these fields in France, within the Agreenium-IAVFF members, while outgoing fellows, already employed by one of these institutes, can spend 6 months to 2 years mobility period at a research laboratory/centre outside France.

As EU co-funded programme, AgreenSkills+ strives to offer its fellows **competitive salary and career development packages**, within first-rate research environments. The programme greatly focuses on the quality of fellowships and their scientific mobility project, the hosting conditions and career prospects.

AgreenSkills+ has a continuous call for applications, with the next selection round deadline on **14 October 2016**.

To apply, visit our website: www.agreenskills.eu

INRA SCIENCE & IMPACT agreenium

UNIVERSITY OF MICHIGAN

## BIOLOGICAL SCIENCES SCHOLARS PROGRAM
### For Junior, Tenure Track Faculty

The University of Michigan Medical School announces recruitment for the Biological Sciences Scholars Program (BSSP) to enhance the institution's strengths in the biological and biomedical research areas.

Now entering its 18th year, the BSSP has led recruitment of outstanding scientists pursuing research in genetics, microbiology, immunology, virology, structural biology, biochemistry, molecular pharmacology, stem cell biology, cancer biology, physiology, cell and developmental biology, bioinformatics, and the neurosciences. The Program seeks individuals with PhD, MD, or MD/PhD degrees, at least two years of postdoctoral research experience, and who have not previously held a tenure-track faculty position. Candidates will show evidence of superlative scientific accomplishment and scholarly promise. Successful candidates will be expected to establish a vigorous, externally-funded research program, and to become leaders in departmental and program activities, including teaching at the medical, graduate, and/or undergraduate levels. Primary departmental affiliation(s) will be determined by the applicant's qualifications and by relevance of the applicant's research program to departmental initiatives and themes. All faculty recruited via the BSSP will be appointed at the Assistant Professor level.

**APPLICATION INSTRUCTIONS:** Please apply to the Scholars Program through the BSSP website at: **http://bssp.med.umich.edu**. A curriculum vitae (including bibliography), a three page research plan, an NIH biosketch, and three original letters of support should all be submitted through the BSSP website. More information about the Scholars Program, instructions for applicants and those submitting letters of recommendation, and how to contact us is located on the BSSP web site: **http://bssp.med.umich.edu**. The deadline for applications is **Friday, September 30, 2016**.

*The University of Michigan is an Affirmative Action/Equal Opportunity Employer.*


# Opportunity awaits.

clusterhiring.ucr.edu

*By* **Zachary S. Wiersma**

# *Wearing my disability with pride*

Snap, cluck, snap! The noises rang out in the conference hall. Several people turned toward me to see what was causing the disturbance. As I walked down the aisle to join my graduate school labmate, I jerked my head to the left and snapped my fingers in my right ear, continuing to draw attention to myself. But when the presentation began, I forgot about how others might see me. I focused solely on the science, even though I clucked my tongue, tightened my right calf, and thrust out my right arm throughout the presentation. None of it was intentional: I suffer from Tourette syndrome (TS), a neurological condition that compels me to perform involuntary repetitive movements and vocalizations known as tics. And I wasn't embarrassed. In fact, after some years of struggling with the potential negative impact of my TS, I now consider it to be an asset.

When I was a child, my tics did not cause any serious problems. But as I got older and thought about my future career and how professional colleagues would perceive me, I worried my TS might impede my dreams. By the time I began college, I was acutely self-conscious about my disorder and worried that it might distract teachers and fellow students or deter them from interacting with me.

These concerns seemed justified when, as a sophomore, I toured a pharmaceutical plant with other students. The research and commercial potential fascinated me, and I spent all morning entranced by the science. But my musing was interrupted when I entered the conference hall for a presentation and heard students and staff members complaining about me under their breath. I noticed a few stares and some furtive pointing. At first, I didn't understand. And then I realized: They had been watching me snap my fingers in my ear and cluck my tongue all day.

I was flustered. Pharmaceutical research and its broader impacts were my focus that day, and I couldn't understand why anyone would pay attention to me instead. I was embarrassed and convinced I had distracted everyone, maybe even causing them to miss out on opportunities. I asked my undergraduate adviser, who was accompanying us, if I should leave or sit in the back of the room. He calmed my concerns with a grace and wisdom I will never forget. "Let them talk," he said. "They'll want to learn who you are."

In that moment, I reshaped my perspective of disabilities in professional settings. I realized that my TS is a powerful tool that can help me connect and communicate with others. I read people better than most because I make a point of looking past their exteriors and any unusual behaviors. At the same time, I am particularly good at observing and analyzing body language.

***"We can all allow our disabilities to be sources of strength."***

I am now a fourth-year Ph.D. candidate. I also work part-time at the university technology transfer office, where I meet with professors to discuss their inventions. Back in college, my TS would have embarrassed and distracted me during these meetings, but now I explain my disorder without shame. I exhibit confidence and self-awareness and don't allow room for judgment. In turn, most professors warm to me quickly and lower their guard.

My experiences at the technology transfer office, coupled with my love for research, have inspired a passion for academia and invention, and I hope to become a professor and an academic entrepreneur. Shifting my perspective about my TS helped me see that I can succeed.

I also plan to mentor and advocate for other scientists with disabilities. I recognize that some face accessibility challenges that cannot be overcome with a change in mindset. But regardless of these barriers, I hope we can all allow our disabilities to be sources of strength. We shouldn't hide our true selves because of fears of blemished careers. I look forward to adding my voice and actions to broader efforts to make science more accessible and inclusive for scientists with disabilities. Together we can all wear our disabilities with pride. ■

*Zachary S. Wiersma is a Ph.D. candidate in chemistry at the University of Illinois, Urbana-Champaign. Send your story to SciCareerEditor@aaas.org.*