New sanctions squeeze Iran's science
p. 961

Fresh fall books
p. 966

Solar greening in the Sahara
p. 1019

Science

$15
7 SEPTEMBER 2018
sciencemag.org

AAAS

LEARNING THE WAY

Social information guides ungulates on seasonal migrations *pp. 972 & 1023*

CONTENTS

962
Is it too costly to save some caribou herds?

7 SEPTEMBER 2018 • VOLUME 361 • ISSUE 6406

Fire destroys priceless specimens at the National Museum of Brazil.

NEWS

IN BRIEF

IN DEPTH

FEATURES

INSIGHTS

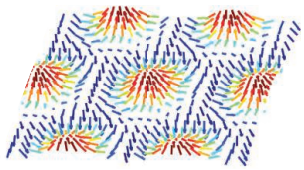
BOOKS *ET AL.*

PERSPECTIVES

POLICY FORUM

LETTERS

CONTENTS



993 Optical skyrmions

7 SEPTEMBER 2018 • VOLUME 361 • ISSUE 6406

RESEARCH

IN BRIEF

REVIEW

RESEARCH ARTICLES

REPORTS

DEPARTMENTS

ON THE COVER

Pronghorn (*Antilocapra americana*) ford the Green River in Wyoming, United States, on a 160-km migration in search of high-quality food. To navigate successfully, ungulates (hooved mammals) must learn when and where to migrate, and reintroduced populations of bighorn sheep and moose demonstrate that this knowledge is passed across generations. Conservation of migration corridors helps these animals maintain this cultural knowledge and sustain abundant populations. See pages 972 and 1023.
Photo: © Joe Riis

SCIENCE (ISSN 0036-8075) is published weekly on Friday, except 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, including subscription (12 months): $165 ($74 allocated to subscription). Domestic institutional subscription (51 issues): $1808; 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 #125488122. 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 each plus shipping and handling; bulk rate on request. **Authorization to reproduce** material for internal or personal use under circumstances not falling within the fair use provisions of the Copyright Act is granted by AAAS to libraries and others who use Copyright Clearance Center (CCC) Pay-Per-Use services provided that $35.00 per article is paid directly to CCC, 222 Rosewood Drive, Danvers, MA 01923. The identification code for *Science* is 0036-8075. *Science* is indexed in the *Reader's Guide to Periodical Literature* and in several specialized indexes.

CREDITS: (GRAPHIC) S. TSESSES *ET AL.*; (PHOTO) IMP

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 Central Support** 866-434-2227 www.aaas.org/membercentral.

INFORMATION FOR AUTHORS See www.sciencemag.org/authors/science-information-authors

Published by AAAS

EDITORIAL

New voices, at last

Science academies are well placed to help strengthen national systems of research and innovation, drawing upon the knowledge of experienced scientists across fields. But what about the voices of early-career scientists who are impassioned to bolster international, interdisciplinary, and intergenerational dialogue, with the goal of making decision-making evidence-based and inclusive? Since 2000, "young" science academies have emerged across the world. The United States has now joined this movement. In June, the U.S. National Academies launched "New Voices in Sciences, Engineering and Medicine," comprising 18 early- to mid-career scholars. They will engage in communicating the evidence base for addressing national and global challenges and help diversify the expertise in the National Academies' advisory activities (www.nationalacademies.org/newvoices). As New Voices convenes its first meeting next week, what should this organization consider?

"New Voices will be warmly welcomed by the global network of Young Academies."

Inspiration can be drawn from abroad. Starting in Germany, there are now some 50 nations that have a Young Academy–like institution. There are also supranational ones such as the vibrant Global Young Academy (www.globalyoungacademy.net). Scientists with disciplinary backgrounds that range from natural sciences to humanities and from medicine to engineering are selected as members on the basis of their research excellence and commitment to society. Typically, they are up to 40 years of age upon entry. Because they serve only a single 4- to 5-year term, they are incentivized to get things done, within a time frame that is feasible for academics who are under career pressure. It also allows for continual renewal of the academy.

Areas of Young Academy activity are policy-making, education, outreach, capacity building, addressing young scientists' career concerns, and promoting open science and collaboration. Examples include involvement in the European Commission's Scientific Advice Mechanism and the United Nations Sustainable Development Agenda, the workshops on Responsible Conduct of Research in Malaysia, and contributions to gender equity and stable funding policies by the Early- and Mid-Career Researcher Forum in Australia. Young Academies' members and alumni have also developed successful cross-disciplinary Science Leadership Programmes in Africa and Southeast Asia. And they support those who face barriers in accessing academia. The Dutch, Scottish, and Global Young Academies, for example, have support programs for early-career scholars who are displaced or are refugees, in partnership with organizations such as Scholars at Risk.

Both of us have long been involved in the Young Academy movement and consider the following elements important for a Young Academy's success: careful and transparent membership-selection processes that value diversity; a size and structure that enable the organization to think and act big, while not being so large that members may lose connection to each other and to the whole (typically 50 to 200 members, with secretarial support; the inaugural 18 members of New Voices may be a good start for further growth); and autonomy for the organization to set its own agenda.

We encourage the first members of New Voices to be bold. With so many major societal issues that call for the involvement of younger generations, New Voices could become an important stakeholder for policy-making. They can also empower other young scientists to contribute to this movement. Strong partners may furthermore be found in, for example, popular culture and media, charitable foundations, industry, and politics. New Voices will be warmly welcomed by the global network of Young Academies. It will hopefully spur the United States to become a major participant in this worldwide effort to spread the benefits of scientific developments, in their broadest sense, to all.

–Eva Alisic and Hans Hilgenkamp

Eva Alisic
is an associate professor and associate director of the Jack Brockhoff Child Health and Wellbeing Program, Melbourne School of Population and Global Health, University of Melbourne, Australia. ealisic@unimelb.edu.au

Hans Hilgenkamp
is a professor of physics at the Faculty of Science and Technology and MESA+ Institute for Nanotechnology, University of Twente, the Netherlands. h.hilgenkamp@utwente.nl

10.1126/science.aav2338

$160,000,000

Donated to Yale University to renovate its Peabody Museum of Natural History, in the largest gift to a U.S. natural history museum, by Yale alumnus Edward Bass, known for also funding the Biosphere 2 ecology project in Arizona.

IN BRIEF

Edited by **Jeffrey Brainard**

BIOANTHROPOLOGY

Germany returns stolen skulls

Researchers acquired this and other skulls of victims killed in a genocide in Namibia in the early 1900s.

The German government returned remains of 25 victims of a colonial-era genocide in Namibia to representatives of that country last week in Berlin. The 19 skulls, as well as bones and skin samples, had been taken to Germany in the early 20th century for "anthropological" research. They belonged to victims of an extermination campaign in which the German colonial army killed an estimated 75,000 Herero and Nama people. The remains, collected from half a dozen cities, are the third set to be transferred since 2011; researchers suspect more are stored in anatomy collections in museums, hospitals, universities, and some private hands, but it is a challenge to identify them. Representatives of the Herero and Nama want reparations from the German government, which in 2015 acknowledged the genocide. No agreement has been reached.

U.K. institute probes bullying

WORKPLACE | The Wellcome Sanger Institute in Hinxton, U.K., is investigating allegations of bullying by some of its management team. *The Guardian* reported last week that 10 members of senior staff had filed complaints claiming they were mistreated, bullied, and pressured to leave the institute on short notice. An institute spokesperson confirmed that an investigation is ongoing. The institute, which specializes in genome sequencing and human health, employs nearly 1000 people; it is the largest scientific institute owned by the Wellcome Trust, the United Kingdom's largest medical charity. In May, Wellcome announced a new antibullying policy with strict sanctions, including a loss of funding, for those found to have violated guidelines.

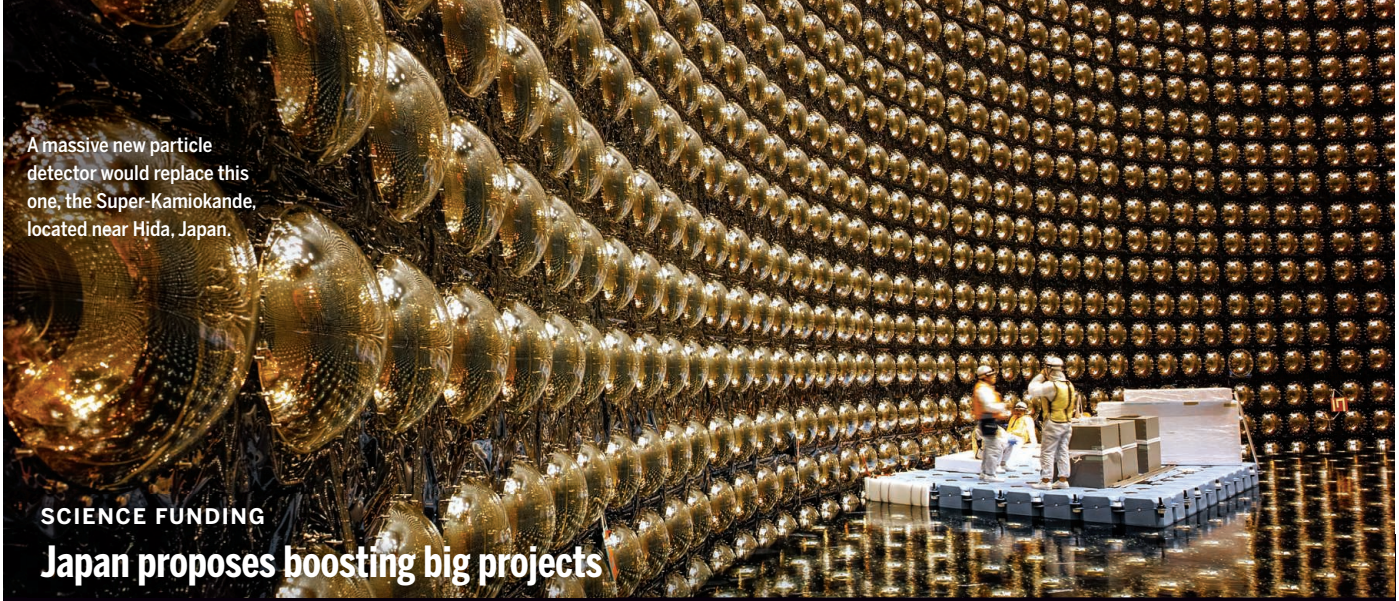
Editors quit journal over quality

PUBLISHING | All 10 senior editors of the open-access journal *Nutrients* resigned last month, alleging pressure by the publisher, the Multidisciplinary Digital Publishing Institute (MDPI) in Basil, Switzerland, to accept mediocre manuscripts. MDPI and other open-access publishers collect fees from authors per published article and thus have a financial incentive not to be too strict. The journal's former editor-in-chief, Jon Buckley of the University of South Australia in Adelaide, says he quit because MDPI wanted him to publish more manuscripts and be less choosy, whereas he focused on maintaining quality; the human clinical nutrition journal's impact factor rose under his leadership. MDPI CEO Franck Vazquez says the company wants to publish all articles that are "good enough" and "useful to researchers," with less emphasis on novelty and impact factor. News of the dispute emerged just as 11 European national funding organizations announced that research they fund should be published only in open-access journals from 2020 onward (see story, p. 957).

Cancer trial conflicts exposed

CLINICAL TRIALS | Oncologists performing clinical tests of cancer drugs received $80 million in undisclosed payments from

PHOTO: CHRISTIAN MANG/REUTERS



A massive new particle detector would replace this one, the Super-Kamiokande, located near Hida, Japan.

PHOTO: C. KAMIOKA OBSERVATORY, INSTITUTE FOR COSMIC RAY RESEARCH, THE UNIVERSITY OF TOKYO

SCIENCE FUNDING

Japan proposes boosting big projects

Japan's main science ministry last week proposed an ambitious budget for basic research that would allow Japan to compete for the world's fastest supercomputer and push ahead with a massive new particle detector. The blueprint by the Ministry of Education (MEXT) represents a 21% increase for fiscal year 2019, to 1.17 trillion yen ($10.54 billion). However, the final figure is certain to be scaled back because government leaders face serious fiscal challenges and want to trim overall discretionary spending. As Japan's population has aged and tax revenues have shrunk, spending on R&D has been flat for most of the past decade, until a 7% increase this year. Next year, big-ticket items funded by MEXT are proposed to get the most generous increases. Funding would more than triple for Japan's next-generation, exascale supercomputer. The budget would also finance a feasibility study for the Hyper-Kamiokande detector, a giant water-filled tank lined with sensors that would pick up the flashes generated when neutrinos from supernovae collide with water molecules. The tank would hold 10 times as much ultrapure water as its predecessor, Super-Kamiokande, vastly increasing the data collected. Meanwhile, funding for research grants to academic groups and individuals would rise 8%.

pharmaceutical companies sponsoring the drugs and trials while they were underway, a study has found. The analysis examined author financial disclosure statements in publications of clinical trial results for cancer drugs approved by the U.S. Food and Drug Administration between January 2016 and August 2017; these were compared against payments made by sponsors to physicians, as recorded in the Centers for Medicare & Medicaid Services's Open Payments database. The payments included research grants, stock in companies, speaking fees, travel reimbursements, and fees to enroll patients or coordinate the trials. Of 1007 oncologist authors, 76.5% received at least one industry payment, and 32% did not fully disclose the payments in the publication, according to the study, published last week in *JAMA Oncology*. The analysis's authors warn that although financial relationships between drug sponsors and physicians can help advance the development of lifesaving drugs, they may also introduce bias into studies of drug effectiveness.

Gender gap seen in peer reviews

PEER REVIEW | All-male teams of peer reviewers are more likely to accept manuscripts with a male senior author than those with a female one, reports an un–peer-reviewed study. But a gender gap in acceptance rates wasn't observed for manuscripts reviewed by mixed-gender teams, according to the study, posted 29 August on the bioRxiv server. Researchers examined 5495 manuscripts submitted to the life sciences journal *eLife* from 2012 to 2017, compiling information about the genders of the journal's review teams and study authors, including the last authors—a position that in the life sciences commonly denotes the most senior researcher. When all of the reviewers were male, manuscripts with a male last author were accepted at a rate 6% higher than manuscripts with a female last author. *eLife*'s review process is unusual because reviewers interact with one another, so it's not clear whether the gender disparity detected by the study occurs in other journals.

Documentary critiques paywalls

PUBLISHING | Jason Schmitt was working at Atlantic Records in the early 2000s when the online site Napster disrupted the music industry by making copyrighted songs freely available. Now, the communications and media researcher at Clarkson University in Potsdam, New York, is pushing for a similar disruption of academic publishing with *Paywall*, a documentary about the open-access movement. "I don't think that it's right that for-profit publishers can make 35%–40% profit margins. The content is provided for them for free by academics," Schmitt, who produced the film, says in an interview. (John Bohannon, a former contributing correspondent for *Science*, was a consultant on the project.) Schmitt says many large publishers refused to go on camera—although representatives from *Science* and *Nature* did—and he is not impressed that several have begun publishing some open-access journals. "Elsevier is as much to open access as McDonald's fast food is to healthy," he says. Schmitt adds that because so many U.S. universities separately negotiate journal subscription fees with publishers, the country is slowing a movement now led by other nations. *Paywall* was scheduled to make its debut this week in Washington, D.C.; it can also be streamed for free online.

CORRECTION

An In Brief item headlined "Carbon plan sinks Australian PM" (31 August, p. 829) erroneously reported that former Australian Prime Minister Malcom Turnbull "refused" to drop a greenhouse gas emissions target to appease members of the country's conservative coalition. He did not refuse, but was still pushed out in a leadership vote.

IN DEPTH

MARINE SCIENCE

U.N. tackles gene prospecting on the high seas

Talks on marine biodiversity pact marked by north-south divide on genetic resources

By **Eli Kintisch**

It's an eye-catching statistic: A single company, the multinational chemical giant BASF, owns nearly half of the patents issued on 13,000 DNA sequences from marine organisms. That number is now helping fuel high-stakes global negotiations on a contentious question: how to fairly regulate the growing exploitation of genes collected in the open ocean, beyond any nation's jurisdiction.

The overarching goal of the talks, which opened this week at the United Nations in New York City, is crafting a new agreement to protect biodiversity in the high seas, which include two-thirds of the ocean. Many of the discussions, which will run until 17 September, are expected to focus on long-standing proposals to establish protected zones where fishing and development would be limited or banned. But the negotiations also aim to replace today's free-for-all scramble for marine genetic resources with a more orderly and perhaps more just regime.

Many developed nations and industry groups are adamant that any new rules should not complicate efforts to discover and patent marine genes that may help create better chemicals, cosmetics, and crops. But many developing nations want rules that will ensure they, too, share in any benefits. Scientists are also watching. A regulatory regime that is too burdensome could have "a negative impact" on scientists engaged in "noncommercial ocean research," warns Robert Blasiak, a marine policy specialist at the Stockholm Resilience Centre.

It is not the first time nations have wrangled over how to share genetic resources. Under another U.N. pact, the 2010 Nagoya Protocol, 105 countries have agreed to rules to prevent so-called biopiracy: the removal of biological resources—such as plant or animal DNA—from a nation's habitats without proper permission or compensation.

Those rules don't apply in international waters, which begin 200 nautical miles from shore and are attracting growing interest from researchers and companies searching for valuable genes. The first patent on DNA from a marine organism was granted in 1988 for a sequence from the European eel, which spends part of its life in freshwater. Since then, more than 300 companies, universities, and others have laid claim to sequences from 862 marine species, a team led by Blasiak reported in June in *Science Advances*. Extremophiles have been especially prized. Genes from worms found in deep-sea hydrothermal vents, for example, encode polymers used in cosmetics. And BASF has patented other worm DNA that the company believes could help improve crop yields. The conglomerate, based in Ludwigshafen, Germany, says it found most of its 5700 sequences in public databases.

"There's nothing illegal about what they did, the science is just moving way faster than the policy," Blasiak says. And the trend has helped catalyze global discussions about whether and how to share any benefits that arise from high seas patents.

It may take years for nations to agree on a marine biodiversity treaty; the talks now underway are just a first round. But an "ideological divide" between developing and developed countries has, so far, "led to stalemate" on how to handle marine genetic resources, says Harriet Harden-Davies, a policy expert at the University of Wollongong in Australia.

Most developing nations want to expand the "common heritage" philosophy embedded in the 1982 United Nations Convention on the Law of the Sea, which declares that resources found on or under the seabed, such as minerals, are the "common heritage of mankind." Applying that principle to genetic resources would promote "solidarity

PHOTO: PASQUALE VASSALLO/GETTY IMAGES

The purple-striped jellyfish is among the hundreds of marine organisms that have had their DNA patented.

in the preservation and conservation of a good we all share," South Africa's negotiating team said in a recent statement. Under such an approach, those who profit from marine genes could, for example, pay into a global fund that would be used to compensate other nations for the use of shared resources, possibly supporting scientific training or conservation.

But developed nations including the United States, Russia, and Japan oppose extending the "common heritage" language, fearing burdensome and unworkable regulations. They argue access to high seas genes should be guaranteed to all nations under the principle of the "freedom of the high seas," also enshrined in the Law of the Sea. That approach essentially amounts to finders keepers, although countries traditionally have balanced unfettered access with other principles, such as the value of conservation, in developing rules for shipping, fishing, and research in international waters.

The European Union and other parties want to sidestep the debate and seek a middle ground. One influential proposal would allow nations to prospect for high seas genes, but require that they publish the sequences they uncover. Companies could also choose to keep sequences private temporarily, in order to be able to patent them, if they contribute to an international fund that would support marine research by poorer nations. "Researchers all around the world should be put all on a level playing field," says Arianna Broggiato, a Brussels-based legal adviser for the consultancy eCoast, who co-authored a paper on the concept this year in *The International Journal of Marine and Coastal Law*.

Many researchers, meanwhile, hope whatever rules emerge don't stifle science. Some complain that the Nagoya Protocol, for example, has sometimes led to burdensome paperwork, complicating even studies aimed at protecting biodiversity. Worse, many say, is a recent proposal to use the pact to regulate the use of not just genetic material, but also publicly available gene sequences (*Science*, 6 July, p. 14).

It's possible that nations ultimately won't reach a new agreement and will maintain the status quo. But if developed nations continue to claim valuable genetic information from the high seas, developing nations could have little incentive to protect those waters, warns marine policy expert Thembile Joyini, a New York City–based adviser to South Africa's government. Speaking as an individual and not representing his government, he recently said, "You want all of us to have the feeling that we own the ocean." ■

SCIENTIFIC PUBLISHING

European funders seek to end reign of paywalled journals

Move aims to accelerate full transition to open access

By **Martin Enserink**

Frustrated with the slow transition toward open access (OA) in scientific publishing, 11 national funding organizations in Europe turned up the pressure this week. As of 2020, the informal group, which jointly spends about €7.6 billion on research annually, will require every paper resulting from research funded by its members to be freely available from the moment of publication. In a 4 September statement, they said they will no longer allow the 6- or 12-month delays that many subscription journals now require before a paper is made OA. The coalition will also ban publication in so-called hybrid journals, which charge subscriptions but also make individual papers OA for an extra fee.

> ***"Research communities just aren't willing to tolerate procrastination anymore."***
>
> **Ralf Schimmer**, Max Planck Digital Library

The move means grantees from these funders—which include the national funding agencies in the United Kingdom, the Netherlands, and France as well as Italy's National Institute for Nuclear Physics—will have to forgo publishing in thousands of journals, including high-profile ones such as *Nature*, *Science*, *Cell*, and *The Lancet*, unless those journals change their business model. "We think this could create a tipping point," says Marc Schiltz, president of Science Europe, the Brussels-based association of science organizations that helped coordinate the plan. "Really the idea was to make a big, decisive step—not to come up with another statement or an expression of intent." (The plan's name, Plan S, is a "working title that essentially stuck," says a Science Europe spokesperson; the S could stand for "speed" or "stop paywalls," they say.)

The plan could cause turmoil in the publishing industry and push it further toward OA, says Ralf Schimmer, head of scientific information provision at the Max Planck Digital Library in Munich, Germany. "This will put increased pressure on publishers and on the consciousness of individual researchers that an ecosystem change is possible," he says. Peter Suber, director of the Harvard Library Office for Scholarly Communication, calls the plan "admirably strong." Many other science funders support OA, but only the Bill & Melinda Gates Foundation applies similarly stringent requirements for "immediate OA," Suber says.

The European Commission and the European Research Council support the plan, although they haven't adopted similar requirements for the research they fund. But in a 4 September statement, EU Commissioner for Research, Science and Innovation Carlos Moedas suggested they may do so in the future; he urged the European Parliament and the European Council to endorse the approach.

Traditional publishers are not pleased. The plan "potentially undermines the whole research publishing system," a spokesperson for Springer Nature, which publishes more than 3000 journals, wrote in an email to *Science*. "Implementing such a plan, in our view, would disrupt scholarly communications, be a disservice to researchers, and impinge academic freedom," adds a spokesperson for AAAS, *Science*'s publisher. "It would also be unsustainable for the *Science* family of journals." The world's biggest academic publisher, Elsevier, declined to comment, referring instead to a statement by the International Association of Scientific, Technical, and Medical Publishers that urged "caution" and said, "Above all, it is vital that researchers have the freedom to publish in the publication outlet of their choice."

Europe has taken the lead in pushing for OA in recent years; EU ministers of research, innovation, trade, and industry announced a 2020 target for making all new research papers freely available at a meeting in Brussels in 2016. But the commission's special envoy for OA, Robert-Jan Smits, says the transition is taking far too long. Smits was the "catalyst" behind the new plan, says Stan Gielen, president of the Netherlands Organization for Scientific Research (NWO) in The Hague.

Under Plan S, authors need to retain the copyright on their papers and publish them under a licensing scheme that allows free re-

use. The funders will cap the fees paid for publication in OA journals at a yet-to-be-determined level. After a transition period, publication in hybrid journals—of which Springer Nature operates more than 1700 and Elsevier more than 1850—will end because such journals have not proved to be the transition model that many were hoping for, Schiltz says. In fact, he adds, "We now pay more" because the author publication fees come on top of the subscription price. (The Springer Nature statement says hybrid journals do "support the transition towards full open access"; under special agreements, they allow 70% of authors in four European countries to make their research available immediately.)

The plan is ambivalent about "green OA," in which researchers or institutions post a freely accessible copy of their paper in an institutional repository, instead of publishing in an OA journal; it only says the importance of such repositories is "acknowledged." That's an "elementary mistake," Suber says, because green OA has its own advantages. Also called self-archiving, it is cheap and easy to scale up, and by allowing researchers to make their work freely available while publishing in a "conventional, venerable" journal, green OA helps young scientists who need the cachet of publishing in top journals, Suber says.

The funders behind the new plan, however, explicitly aim to reduce the allure of marquee journals. In a preamble to Plan S, they pledge to help "fundamentally revise the incentive and reward system of science," for instance by following the 2013 San Francisco Declaration on Research Assessment, which advocates abandoning simple metrics such as the journal impact factor when evaluating a scientist's performance.

Schiltz rejects the claim that the plan infringes on academic freedom. Authors still have plenty of journals to choose from, he says, and funders are entitled to say how their money is spent. "The greater good of a well-functioning science system is more important than the right of individual researchers to decide where to submit their papers," he says.

The funders say they will monitor researchers' publication practices and sanction noncompliance. For example, Gielen says NWO will check a certain percentage of the papers it has funded and could punish researchers who have not followed the new rules by asking for its money back or temporarily banning them from applying for funding.

Many of Science Europe's 18 other funders are likely to come on board in the weeks and months ahead, Schiltz predicts. "We felt this was now strong enough to go public," he says, in part because the 11 current participants jointly represent more than half of the funding stream that Science Europe's members control.

Plan S comes at a time when academic institutions in several European countries, seeking to make more papers OA, are in tough negotiations with academic publishers. Elsevier recently cut off access to its journals in Germany and Sweden after consortia of labs and universities in those countries refused to back down. Plan S will further increase the pressure, Schimmer says. "There has been enough nice language and waiting and hoping and saying please," he says. "Research communities just aren't willing to tolerate procrastination anymore." ■

Robert-Jan Smits, the European Commission's special envoy for open access, was the "catalyst" behind Plan S.

PHOTO: AP PHOTO/LUIGI MISTRULLI/SIPA

PSYCHOLOGY

'Rapid onset' of transgender identity ignites storm

Critics charge a study is biased, but others say politics is inhibiting science

By **Meredith Wadman**

A study describing "rapid onset gender dysphoria" (ROGD) in teens and young adults—a sudden unease with the gender they were assigned at birth—has infuriated transgender activists while sparking a debate about academic freedom. Critics of the paper, published last month in *PLOS ONE* by physician-scientist Lisa Littman of Brown University, call it a flawed study that reflects an antitransgender agenda, in part because it suggests some cases may be the result of "social contagion." Brown and the journal have both distanced themselves from the paper, drawing charges that they surrendered to political pressure.

The study remains freely available, but last week, *PLOS ONE* announced it is conducting a postpublication investigation of its methodology and analysis. "This is not about suppressing academic freedom or scientific research. This is about the scientific content itself—whether there is anything that needs to be looked into or corrected," *PLOS ONE* Editor-in-Chief Joerg Heber in San Francisco, California, told *Science* in an interview.

Also last week, Brown officials removed the university's press release highlighting the paper from its website. Bess Marcus, dean of Brown's School of Public Health, wrote in an open letter that the university acted "in light of questions raised about research design and data collection related to the study." She added that people in the Brown community have raised concerns that the study's conclusions "could be used to discredit efforts to support transgender youth and invalidate the perspectives of members of the transgender community."

Brown's move prompted Jeffrey Flier, a former dean of Harvard Medical School in Boston and a professor of medicine there,

to say in a tweet: "This is a sad day for @BrownUniversity, and an indictment of the integrity of their academic and administrative leadership." In an interview, Flier called elements of Marcus's statement "anti-intellectual" and "completely antithetical to academic freedom," and said he found it "horrifying" that Brown failed to defend Littman. A petition urging Brown and *PLOS ONE* "to resist ideologically-based attempts to squelch controversial research evidence" had garnered nearly 3900 signatures by early this week.

The controversy comes after several years of rapid growth in the number of adolescents being referred to clinics specializing in gender dysphoria in North America and Europe. For instance, a paper published in April in the *Archives of Sexual Behavior* analyzed 2009–16 data from a U.K. specialist service that is the largest in the world. The study described recent, dramatic growth in both total adolescent referrals and the proportion of those patients who were designated female at birth (see graph, right). In the past, the majority of patients at such clinics had been designated male at birth. The authors wrote that their findings "reflect a general trend of inversion in sex ratios" in adolescents seeking treatment in several developed countries. They speculated that causes might include that "coming out ... may be easier for birth-assigned females than it is for birth-assigned males" as awareness of transgender identity grows. But, "It is not possible to say with any confidence why" the sex inversion is happening, says Polly Carmichael, the paper's senior author and director of the Gender Identity Development Service at The Tavistock and Portman NHS Foundation Trust in London.

Michael Bailey, an academic psychologist who studies sexual orientation and gender dysphoria at Northwestern University in Evanston, Illinois, says his colleagues who treat gender dysphoria "all tell me that their primary group these days are adolescent females who were not known to be gender dysphoric [in childhood]. ... This kind of case virtually never happened until recently—even a decade ago you didn't see them. I don't know what else to call this but an epidemic."

In 2016, spurred by accounts of sudden transitions among young people, Littman surveyed parents she recruited from three websites where she had read such descriptions by parents: 4thWaveNow, Transgender Trend, and YouthTransCritical Professionals. The first two are gathering places for parents concerned by their children's exploration of a transgender identity. (The third website is closed to nonmembers.)

According to 256 parents who responded to the 90-question survey, none of their children—83% of whom were designated female at birth—had symptoms that matched the professionally defined diagnosis of gender dysphoria during childhood. The finding suggests "that not all [young people] presenting at these vulnerable ages are correct in their self-assessment of the cause of their symptoms," Littman wrote. She suggested some young people may be seeking gender transition to escape other emotional difficulties.

But transgender activists furiously dispute the existence of ROGD, and Littman's description of it, which is the first in the literature. They argue that what may seem a "rapid onset" to parents is likely the result of a lengthy internal process in children. "What's 'rapid' about ROGD is parents' sudden awareness and assessment of their child's gender dysphoria," the Oakland, California–based transgender writer and former developmental biologist Julia Serano wrote in a critical essay last month. She argues that Littman's paper provides no evidence for the existence of ROGD. She added in an interview that others have already embraced ROGD for ideological reasons—"to do an end run around existing trans health practices that advocate for supporting and affirming trans kids."

The most explosive of Littman's findings may be that among the young people reported on, more than one-third had friendship groups in which 50% or more of the youths began to identify as transgender in a similar time frame. This, Littman writes, was more than 70 times the expected prevalence of transgender identity in young adults. She hypothesizes that "social contagion" may be a key driver of some cases of the purportedly rapid onset dysphoria. To trans activists and some clinicians, such a suggestion denies the inner experience of transgender youths and risks stigmatizing and further isolating them from their peers and supportive resources.

Critics also assailed Littman for failing to recruit participants from other websites supportive of transgender youth and for not interviewing such youths themselves. Littman defended her choice of sites, writing in an email to *Science* that in order to find cases of ROGD, she targeted the only three sites where she had seen parents discussing something like it.

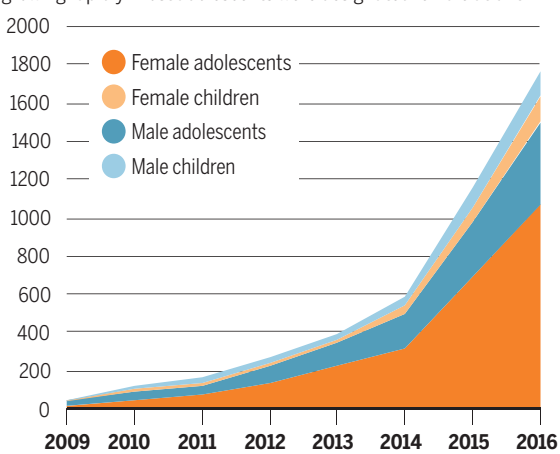
"I would have rejected this manuscript outright for its methodological flaws and also its bias," says Diane Ehrensaft, director of mental health at the Child and Adolescent Gender Center Clinic at the University of California, San Francisco's Benioff Children's Hospital; she treats transgender young people as a clinical psychologist and has reviewed scientific papers for journals. The paper's implication that gender exploration "is simply a fad whipped up by peer influence" should not be taken as authentic, she argues. "It negates the experience of many transgender youth."

But Ray Blanchard, a professor of psychiatry at the University of Toronto in Canada who worked for 15 years in a gender identity clinic that screened candidates for sex reassignment surgery, says the paper points to a clear phenomenon. "Many clinicians in North America and elsewhere have been seeing such patients," Blanchard, who worked with adults, wrote in an email, "and it has been speculated that this subgroup is one reason for the predominance of adolescent females now being seen in North America and elsewhere." Blanchard added, "No one can deny the clinical reality" of a new subgroup of adolescents, mainly female, who experience gender dysphoria after showing no behavioral signs of it during childhood.

In the study, Littman acknowledged its limitations. "Like all first descriptive studies, additional studies will be needed to replicate the findings," she wrote. She told *Science* that in upcoming research she plans to recruit parent-teen pairs in cases where the teenager experienced ROGD that later resolved. ■

Transgender identities on the rise

The number of young people referred to a U.K. transgender service is growing rapidly. Most adolescents were designated female at birth.



GRAPHIC: N. DE GRAAF *ET AL.*, *ARCHIVES OF SEXUAL BEHAVIOR* 47:5, 1301, (2018). ADAPTED BY A. CUADRA/*SCIENCE*

The 19th century museum, seen the day after the fire, had dangerous wiring and little fire protection.

PHOTO: MAURO PIMENTEL/AFP/GETTY IMAGES

CULTURAL HERITAGE

In a 'foretold tragedy,' fire consumes Brazil museum

Years of underfunding are blamed for lack of fire safety

By **Herton Escobar**

Like most Brazilians, paleontologist Sandro Scheffer is a regular coffee drinker. But he and many colleagues at the National Museum of Brazil in Rio de Janeiro preferred not to have a coffee machine at the office. They worried that the museum's old wiring couldn't take the load. "We were always afraid the building would catch on fire."

On the night of 2 September, their fears were realized when a fierce blaze gutted the historical yellow structure. The cause of the fire, which left only the exterior walls standing, was not known at press time, and no one was hurt. But in a matter of hours, 2 centuries of Brazil's most prized scientific and cultural heritage were erased. Founded in 1818 and housed since 1892 in what had been a palace for Brazil's Portuguese rulers, the museum was the country's oldest scientific institution, with massive archives and archaeological and natural history collections numbering perhaps 20 million items.

Parts of the vertebrate and botanical collections were stored in a separate building and escaped the blaze. Almost everything else may have been lost; Cristiana Serejo, a vice-director of the museum, initially estimated that 90% of the collections were destroyed. "It's an irreparable loss, not only for Brazilian science, but for the world. The building can be reconstructed, restored, and everything else, but the collections can never be replaced. Two centuries of science and culture are lost forever," says Sergio Alex Kugland de Azevedo, a paleozoologist and former director of the museum.

Among its most prized items were the 11,500-year-old skull called Luzia, one of the oldest human remains in the Americas; a 5-ton meteorite called Bendegó, the largest found in Brazil; several type specimens of South American dinosaurs and pterosaurs—the specialty of the museum's recently appointed director, paleontologist Alexander Kellner—and a world-famous insect collection with more than 5 million specimens, as well as mummies, sarcophagi, and other ancient Egyptian artifacts. At press time, the meteorite was known to have survived; the fate of other artifacts was not known.

In recent years budget woes had plagued the museum, managed by the Federal University of Rio de Janeiro (UFRJ) with dwindling funds from the federal government. Staff had warned as early as 2004 of dangerous wiring and a lack of fire protection. The building's interior was mostly wood, and it had no sprinkler system; limited water was available from fire hydrants when firefighters arrived. The museum's 200-year anniversary in June was celebrated with the announcement of a 21.7 million real ($5 million) grant from the Brazilian Development Bank, meant to fund upgrades to the museum's infrastructure—including fire safety systems. But it came too late.

"It was a foretold tragedy," says herpetologist Hussam Zaher of the University of São Paulo's (USP's) Museum of Zoology in São Paulo, Brazil, who started his scientific career at the National Museum. "It was just a matter of time," adds anthropologist Walter Neves, a retired professor at USP, who described the Luzia skull in the late 1990s and proposed, based on its unusual features, that it represented a first wave of migration to the Americas. "The museum was completely abandoned, left to rot by the disdain and carelessness of public authorities. I am in complete grief," he says.

Along with Luzia, the museum's archaeological material included the oldest skeletons found in Brazil, excavated in the Lagoa Santa region; the country's largest collection of skeletons from the coastal shell midden tradition known as Sambaqui; and a rich variety of precolonial Amazonian material culture and Andean pottery, "among many other treasures that are now forever lost," says André Strauss, a professor at USP's Museum of Archaeology and Ethnology.

UFRJ physicist Luiz Davidovich, president of the Brazilian Academy of Sciences, says the disaster was something "to be mourned officially with flags at half mast," so that everybody would know that "the National Museum is dead." The fire is another blow on top of recent drastic budget cuts, he adds. "It's another sad chapter in the dismantling of Brazilian science—one that affects not only the future of the country, but also its memory."

Paleontologist Stephen Brusatte at The University of Edinburgh, who has worked with the museum's fossil collections, says the tragedy would be equivalent to the American Museum of Natural History in New York City or the British Museum in London going up in flames. "I can't remember in my lifetime a major national museum like this being destroyed to such a degree," he says. "It makes me inordinately sad to think of those millions of specimens and exhibits, the product of 200 years of collection and the life's work of so many hundreds of scientists and explorers, just going up in flame and turning to dust. It makes me want to cry."

Government officials promised emergency funds to reconstruct the museum as quickly as possible. ■

Herton Escobar is a journalist in São Paulo, Brazil. With reporting by Gretchen Vogel.

INTERNATIONAL RELATIONS

Renewed sanctions strangle science in Iran

Brain drain feared amid economic woes and fraying international collaborations

By **Richard Stone**

To explore the genetic diversity of Iran's desert plants, Hossein Akhani and his colleagues used to send DNA samples to a company in Seoul, which provided fast and reliable sequencing. But a few weeks ago, the University of Tehran biologist says, he received a letter from the company explaining that the South Korean government had advised the firm "not to deal with Iran." The reason: the U.S. withdrawal from the 2015 nuclear deal, a multilateral agreement in which Iran agreed to freeze key nuclear activities in return for relief from international economic sanctions.

On 8 August, citing "the full range of threats posed by Iran," U.S. President Donald Trump issued an executive order reimposing U.S. sanctions "as expeditiously as possible." To put pressure on Europe and others still in the pact, the Trump administration also vowed to penalize foreign entities with U.S. interests that continue to trade with Iran. The U.S. withdrawal and the uncertainty that preceded it have shaken the Iranian economy and left the country increasingly isolated. With funds scarce and international ties fraying, "the situation for science is getting worse and worse," Akhani says.

"Western countries have tried to isolate Iran for almost 40 years," says Navid Madani, an Iranian expat biochemist who studies HIV at the Dana-Farber Cancer Institute and Harvard Medical School in Boston. Over the past decade, as sanctions imposed over Iran's nuclear program ratcheted up, Iranian scientists found creative ways to persevere (*Science*, 4 September 2015, p. 1038). But the latest developments are inflicting heavy damage to both research programs and morale, says Hamid Gourabi, a geneticist at the Royan Institute in Tehran. "We're facing a devastating condition for our research centers and universities."

Iran's currency, the rial, has lost more than half its value since January, dealing a heavy blow to science. At the Royan, a global player in stem cell biology and reproductive medicine, Iranian government funding now covers little more than salaries and food and energy subsidies, Gourabi says, forcing the institute to rely more heavily on fees from patients to buy equipment and reagents. However, fewer Iranians these days can afford treatment at the institute. And because of the sanctions, equipment is often available only on the black market, at high prices, Gourabi says.

Meanwhile, several foreign-sponsored clinical trials with Iranian partners have been stopped and "many others are in danger of being suspended," says Ehsan Shamsi Gooshki, a medical ethicist at the Ministry of Health and Medical Education in Tehran. Other collaborations have been scaled back. For example, an initiative to reduce urban health inequalities run by the Wellcome Trust, the London-based biomedical charity, had sought to include Tehran as one of six pilot cities. "We couldn't find any bank willing to transfer funds" to Tehran, says project leader Majid Ezzati, an environmental health scientist at Imperial College London. The project intends to keep Iranian partners "intellectually involved," he says.

Hossein Akhani can no longer send plant samples abroad for sequencing.

PHOTO: EBRAHIM MIRMALEK

Researchers at the Tehran University of Medical Sciences (TUMS) got similar bad news recently. They were slated to team up with a group at the Icahn School of Medicine at Mount Sinai in New York City for a U.S. National Institutes of Health–funded study on cardiovascular disease patterns in Iran's Golestan province. But Mount Sinai was unable to transfer the Iranian share of the grant to Tehran. (Mount Sinai declined to comment.) "In spite of this we are trying to continue the collaboration based on our own funding," says project partner Reza Malekzadeh, director of TUMS's Digestive Disease Research Institute.

Even a major U.S. initiative that had weathered earlier tensions has felt the chill. Since 1990, the National Academies of Sciences, Engineering, and Medicine in Washington, D.C., working with Iranian counterparts, has organized workshops and exchanges in areas such as seismology, water management, and air pollution, involving 1500 scientists from the two countries. That program has been on hold for almost a year. "We're waiting for further foreign policy developments," says its director, Glenn Schweitzer. One uncertainty, sources say, is whether U.S. scientists could obtain limited exemptions from sanctions to collaborate with Iranian colleagues.

Meanwhile, the Trump administration's 2017 travel ban bars most Iranian citizens from entering the United States. And scientists from Europe and other countries are less likely to visit Iran because of a U.S. regulation that predates Trump's presidency. Since early 2016, people from those countries have been disqualified from the U.S. visa waiver program, which eases entry to the United States, if they have visited Iran, Iraq, Libya, Somalia, Sudan, Syria, or Yemen. That requirement discourages scientists from collaborating on projects in Iran for fear of rejection when they apply for a U.S. visa. "This has become one of the worst issues," Gooshki says.

Madani, an Iranian-American who travels once or twice a year to Iran for research visits, says she does not intend to allow the dismal political atmosphere to derail her efforts. "We have to push against the sanctions," she says. Her latest collaboration with Iranian colleagues is a search for candidate anti-HIV compounds in Iranian herbs. With no express mail service that delivers to the United States now operating in Iran, she recently had to fly to Tehran just to pick up the samples for testing in her lab.

But with few Western scientists following Madani's lead, the outlook for science in Iran remains bleak. If things don't improve soon, Gourabi and others fear an exodus of scientific talent. "We are at the beginning of a crisis," he says. ■

Richard Stone is the senior science editor at Tangled Bank Studios in Chevy Chase, Maryland.

SHOULD IT BE SAVED?

Proposals to focus resources on some endangered species and let others go extinct are stirring fierce debate

By **Warren Cornwall**, *in Canada's Lake Revelstoke valley*

Even with round-the-clock attention and hand-gathered lichen for food, caribou calf 15 was doomed once it stepped from a pen here surrounded by 4-meter-high fences. That was clear as wildlife biologist Rob Serrouya crouched in the shadows of a grove of stunted yew trees a kilometer away from the pen and picked up the brown, nylon radio-tracking collar that once encircled the calf's neck.

Serrouya pawed through fallen needles on the forest floor, searching for clues. He lifted a finger to his mouth, tasting for the rusty flavor of blood. The researcher quickly uncovered four tiny shards of bone, still moist and pink. Those, and another telltale bit of bone in fresh bear scat, outlined the story.

A few days earlier, calf 15—3 weeks old and no bigger than a goat—had left the enclosure where it was born. Bound for meadows high in the Monashee Mountains, it faced a climb through a logged forest thick with alders and willows. The gauntlet was formidable to the calf, but an ideal hunting ground for a bear—a black bear, judging by the scat. And so the long struggle to save the Columbia North caribou herd, now down to about 150 animals, suffered another setback. "Major bummer," Serrouya said, gazing at the collar.

Preserving Canada's woodland caribou (*Rangifer tarandus*) is not for the easily discouraged—or for cheapskates. More than half of the herds, found in the mountains of western Canada and in the boreal forests across northern Canada, are in decline. In part, that's because many of the animals live in valuable old-growth timber or atop natural gas and oil, creating a strong temptation for people to encroach on caribou habitat.

The animals' plight has prompted a string of expensive, elaborate rescue operations. Here, for example, conservationists have spent nearly $2 million over the past 5 years to capture up to 20 pregnant Columbia North females each winter and helicopter them to the 9.4-hectare pen, where an electric fence protects them from predators. Managers release the newborn calves once they are old enough to have a better chance of survival.

A woodland caribou skirts a clear-cut forest in British Columbia in Canada. Human impacts have driven many herds to the brink of extinction.

PHOTO: © DAVID MOSKOWITZ

Such desperate measures have had limited results. Although they have helped pause the downward slide of the herd, the population isn't growing. That has made it hard to avoid asking a painful question about the caribou—one that also applies to the more than 26,000 other species around the world that are threatened with extinction. Should people really try to save every population or species, or would letting some vanish—and focusing scarce resources on the ones that stand a better chance—be smarter?

Faced with a gulf between the species in need and the available resources, some scientists are pushing an approach that combines the cold-blooded eye of an accountant with the ruthless decisiveness of a battlefield surgeon. To do the greatest good, they argue, governments need to consider shifting resources from endangered species and populations that are getting too much attention to those not getting enough. That could mean resolving not to spend money on some species for which the chance of success appears low, such as the vaquita, an adorable small porpoise now down to fewer than 30 animals in Mexico's Gulf of California.

It's a controversial idea condemned by some scientists, who argue it is just oil on the already slippery slope to extinction. But government officials are showing interest. In New Zealand and Australia, they have already incorporated the approach into spending decisions. Canada and the United States are considering a similar move.

Even Serrouya, who works for the University of Alberta in Edmonton, Canada, and has spent much of his career studying the nation's woodland caribou, concedes that giving up on some herds is worth considering. But, "It's hard for me to swallow," he admits. "It's a hard one."

THE TERM "TRIAGE"—from the French verb *trier*, meaning to sort—was born on the battlefields of Napoleonic Europe. Faced with a flood of wounded soldiers, French military doctors conceived of a system to decide who got medical attention and who was too far gone. The idea reached conservation biology as early as the 1980s. But in recent years it has moved from scientific journals to the halls of policymakers, thanks in part to an Australian mathematician and conservation scientist, Hugh Possingham.

In the mid-1990s, the former Rhodes scholar was at the University of Adelaide in Australia, teaching courses about using math to optimize economic choices. A devoted conservationist who grew up bird watching with his father, Possingham wondered whether that approach could be retooled to help save species.

Over the following decade, Possingham and others worked to create formulas that could point to the most efficient way to spend money on species preservation. They tried to quantify answers to key questions: What will species restoration projects cost? How likely are they to succeed? How distinct and important is each species? What actions will benefit multiple species or entire ecosystems, bringing the biggest bang for the buck?

Possingham, who last year became chief scientist for the nonprofit Nature Conservancy, headquartered in Arlington, Virginia, is fond of saying people engage in triage all the time. They weigh the available resources—such as time or money—and pick "the best things to do. That's all that everybody does all day, every day," says Possingham, who splits his time between Australia and the United States.

But it's a change from how governments often act. Today, conservation spending is influenced by a complex array of factors, including how close a species is to extinction and the pressure brought by lawsuits, lobbying, and media coverage. The result, Possingham and others argue, is that money is often poured into costly long shots or charismatic organisms, whereas species that could be secured for a relatively low cost go wanting (see graphic, above, and sidebar, p. 965).

A dozen years ago, New Zealand became the first nation to test Possingham's approach. A nation filled with unique species, some 3000 of them at risk, the country is a poster child for the extinction crisis. But New Zealand had no clear process for setting conservation spending priorities, recalls Richard Maloney, a senior scientist in the country's Department of Conservation in Christchurch.

In a bid to do better, officials asked Possingham to help craft a plan for spending roughly $20 million per year. The result was a list of 100 top-priority species, developed using a formula that balanced costs and benefits. In general, highly threatened species unique to New Zealand and ended up at the top of the list. But it also included representatives from a variety of species and took into account the cost and likelihood of success. Before that process, the government was working to recover 130 species. Now, more than 300 are getting attention, Possingham says. In Australia, the state of New South Wales followed suit, and advocates of the strategy say it helped persuade officials there to spend another $100 million over 5 years on conservation.

But for every species or population at the top of such lists, one is at the bottom. And that can lead to agonizing choices.

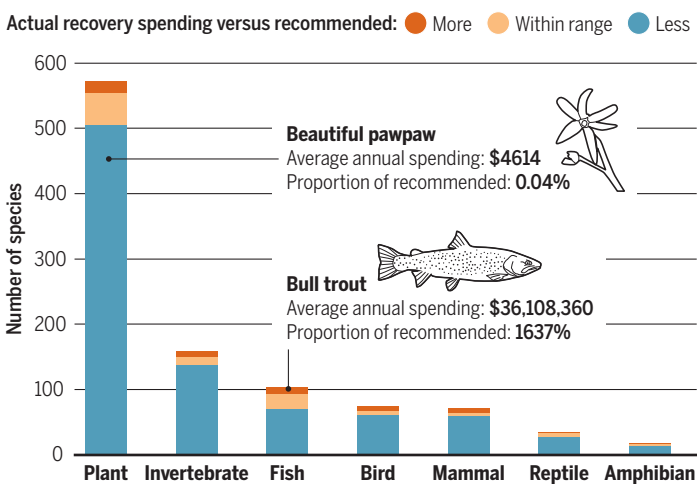
CANADA'S WOODLAND CARIBOU are a symbol of Canadian culture and a keystone for many First Nation peoples. But when humans move in, things don't go well for the animals. Forests cleared for logging, drilling, mining, or roads draw deer and moose that feed on the brush that grows back. The prey, in turn, attracts wolves and mountain lions. Caribou become collateral damage.

Serrouya has watched the results play out near his home in Revelstoke, a small town where workers from the local lumber mill rub shoulders with spandex-clad mountain bikers. Black-haired and stocky, Serrouya evokes the black bears that frequent the craggy, surrounding mountains. Since the late 1990s, he has traversed those slopes on foot and on skis, documenting the lives of caribou and other wildlife. He has seen one nearby caribou herd plummet from 120 animals to just four and another drop from 36 to fewer than 15. About 30 kilometers north of Revelstoke, the Columbia North herd—the area's largest—appeared headed toward a similar fate, falling from 210 animals in the mid-1990s to about 130 by 2004.

To prop up the herd, conservationists have tried a little of everything. Officials banned

Misdirected spending?

The bull trout and a handful of other organisms protected by the U.S. Endangered Species Act get more money for recovery than recommended. But a 2016 study concluded that the vast majority, including the beautiful pawpaw plant, get far less.



GRAPHIC: J. YOU/*SCIENCE*

To save some caribou herds, researchers capture pregnant females and move them to pens to protect the mothers and newborn calves from predators.

logging on roughly half the forest. Hunters shot more moose, hoping to steer wolves elsewhere. Sharpshooters targeted wolves, killing 27 over the past 2 years alone. In 2014, a local nonprofit built the rearing pen.

But saving the caribou might mean keeping the species on life support for decades, Serrouya says. During a drive this summer through the herd's mountainous territory, he pointed to a stripe of light green beside the dark green old-growth forests; it looked as though someone had run a giant lawn mower across the landscape. That logged area is "still producing moose food," Serrouya said, and that means more wolves. He estimates the logged forest could take 20 years to grow back enough to make the area safer for caribou.

The dead calf he found later that day was the latest in a string of bad news. Three cows and two calves had died in the pen over the previous month, prompting fears that a heat wave was killing animals stuck at low elevation. "Under current conditions, even with no worsening habitat," Serrouya predicts the Columbia North herd is "on a trajectory to extinction" unless it gets major help. Farther south, two other herds are already on the brink, having dwindled to tiny single-sex groups—three females in one herd and four males in the other.

Caribou have it bad in Alberta as well. The neighboring province is the heart of the nation's oil and gas industry and is home to 12 caribou herds. Seven of those are in decline, according to a 2017 government report; another three have fewer than 100 animals. There, too, the government is taking extreme measures, shooting and poisoning more than 900 wolves since 2005. And it's considering building a massive enclosure to shield some of the 80 animals left in one herd, at a cost of as much as $15 million over 10 years.

Such schemes reflect the federal government's commitment, under Canada's Species at Risk Act, to save all the nation's caribou herds. But when Mark Hebblewhite, a caribou biologist at the University of Montana in Missoula, looks at how maps of woodland caribou habitat overlap with Alberta's oil and gas deposits, his response is: Get real. Hebblewhite doubts the government will ever summon the will to impose the development restrictions necessary to save all herds. He points to a 2010 study indicating that such restrictions could mean forgoing extraction of oil, gas, and timber worth more than $125 billion in Alberta alone.

Instead of focusing on the most feeble herds, Canada should instead protect habitat in key areas where caribou populations still stand a good chance, he argued in a 2017 *Biological Conservation* paper. "We've prioritized the most screwed populations," Hebblewhite says. "All I'm saying is that we prioritize the winners."

That idea makes biologist Alana Westwood uncomfortable. A Vancouver, Canada–based scientist with the Yellowstone to Yukon Conservation Initiative, she acknowledges that money should be spent where it's most effective. But she fears that openly giving up on some herds cuts at the heart of endangered species laws. "It's really an existential question," she says. If Canada isn't willing to take the necessary steps, she suggests officials rename its law "the 'recover species that are most easy to accommodate under business as usual act.'"

Paths cut for oil and gas surveys expose caribou to wolf attacks.

THE CARIBOU DEBATE underscores one of the major difficulties of the triage approach. It's a lot easier—and publicly acceptable—to say which species or populations should get help than to announce which should be abandoned. That's one reason why scientists working in this field often prefer the term "prioritization" to triage. It's two sides of the same coin, Possingham says, but the feeling is very different. Politicians

PHOTOS: (TOP TO BOTTOM) © ISABELLE GROC; © ANTJE BOHM

don't "tell you what they're not doing," he notes. "They would tell you what they are doing. Because politicians are not stupid."

Canadian officials, however, are starting to ask that once-unthinkable question: Is it worth funding efforts to save every species in need? In 2014, British Columbia officials asked experts to examine how it should spend money to conserve species in the southeastern part of the province. Using an approach similar to Possingham's, they estimated how 60 species would benefit from a menu of actions over 2 decades and how much it would cost. The caribou appeared to be a poor bet, with a less than 50% chance of recovering. A draft of their report dryly notes that "reallocating resources to species and ecosystems that will benefit from investment may need to be considered." In a sign of the sensitivity surrounding the issue, however, the provincial government never released the final analysis.

Nonetheless, some form of prioritization appears to be gaining support in Canada. In British Columbia, prominent conservation scientists are lobbying the provincial government to include a systematic way to rank the effectiveness of species recovery actions in environmental legislation it is crafting. At the federal level, officials in Canada's environment ministry are weighing a similar recommendation. In the United States, academics and federal wildlife officials are discussing ways to use those methods to rethink spending allocations for endangered species.

Such moves have alarmed critics—both scientists and environmental groups—who warn that the triage strategy is at best politically naïve and at worst a sneaky way to undermine species protection. "It's an easy way out for managers who don't have the balls to make tough decisions, and therefore we lose species after species," says Stuart Pimm, a conservation scientist at Duke University in Durham, North Carolina, who has often sparred with Possingham in public forums. One problem, he argues, is that giving up on a species also means abandoning a potent tool to rally the public and the courts. Sometimes charismatic animals such as California condors or polar bears can help build political support for saving endangered species or habitats more broadly.

And some species stand for entire ecosystems, Pimm adds. Consider the Cape Sable seaside sparrow. The innocuous songbird lives in Florida's Everglades, where water diversions threaten its marsh habitat. The species might not rank high in a triage system—in part because other populations of related seaside sparrows exist. But because of how the U.S. Endangered Species Act is structured, efforts to protect the sparrow have required policymakers to reallocate water, benefiting the entire ecosystem. "I'm afraid we have to make more complicated decisions than the simple recipes that Hugh comes up with," Pimm says.

Possingham concedes that triage is not suited to every situation. Europe, for example, has wealthy countries and few native endangered species, which makes saving them all realistic. And sometimes a species is so culturally important that it gets special treatment. New Zealand, for instance, has departed from its triage system to give priority to protecting 50 cherished species, including five species of kiwi birds, the nation's mascot.

But Possingham says opponents aren't acknowledging the nuance and power of a well-thought-out triage approach. One advantage, he says, is that using triage would make allocation decisions more transparent. Tight budgets already are forcing agencies in the United States, Canada, and elsewhere to ignore some teetering species, he argues—they just won't admit it. "I don't understand the con argument," Possingham says. "Why would you want to be inefficient?"

IT'S TOO EARLY TO SAY what a new emphasis on penny pinching might mean for the Columbia North herd. Ask Serrouya about triage for caribou, and he'll agree that it might be too late for some herds, especially where the forests are essentially gone.

But he is not ready to write off the herd he has monitored for so long. The Columbia North herd just needs to hold on—with more than a little help from humans—until the forest grows back and the moose and wolves move on. "There's hope, as long as we can hang on to these guys," he says. "Gotta have hope in this business." ■

How triage became a dirty word

In 2016, Leah Gerber, a conservation biologist at Arizona State University in Tempe, published a paper in the *Proceedings of the National Academy of Sciences* (*PNAS*) that argued the U.S. government could get a better return on its investment in saving endangered species by moving money from some "costly yet futile" recovery efforts to others that are underfunded. She called the approach triage.

Today, after enduring attacks from critics who charge Gerber's work is undermining support for species protection, she is far more guarded about using the "T" word, preferring to talk about setting priorities. "Things that have been said about me, I don't want to repeat," she says.

With a White House openly hostile to environmental regulations and some lawmakers clamoring to rewrite the federal Endangered Species Act (ESA), Gerber's work has pulled her into a tense fray. She has been working with officials at the U.S. Fish and Wildlife Service, as well as other researchers, to devise a new way of allocating the agency's roughly $90 million annual budget for recovering more than 1500 endangered species. The team has devised a digital tool, to be formally unveiled later this year, that will allow managers and others to compare the outcomes of different allocation scenarios. Users might put a priority on saving the direst cases, for instance, or protecting the widest diversity of species.

The effort draws, in part, from that 2016 *PNAS* paper. That study found that to meet the funding estimates included in 1125 recovery plans, the federal government would need about $1.2 billion annually—four times its actual spending. Gerber also calculated that the United States spends $17 million a year more than what the plans recommend on 50 recovery efforts that are failing to meet goals—including those for the northern spotted owl and gopher tortoise. Shifting that "surplus" to "grossly underfunded" recovery efforts, she found, could benefit 182 species, many of them obscure plants, mussels, and insects suffering "injurious neglect."

Some advocates worry the approach could provide cover for attacks on the ESA. Beware the "peril of triage thinking," Kierán Suckling, head of the Center for Biological Diversity (CBD) in Tucson, Arizona, warned in a tweet. In any case, the law wouldn't allow some of the spending shifts envisioned by Gerber, says CBD's Noah Greenwald, who is based in Portland, Oregon. Agencies are often required to spend funds on specific species to compensate for some damage, such as damming a river, he notes. And focusing on triage only distracts from the bigger problem: Recovery spending is "paltry. … We just need to spend a lot more," he says.

Gerber doesn't disagree. But she argues allocation modeling helps reveal the true consequences of current budgeting. "By not funding everything we're making choices," she says. But, "We're just not being transparent about what the choices are." —*W.C.*

AUTUMN BOOKS

New books, fresh for fall

From space weapons to mind reading, the books on this year's list tell tales of scientific transformation, balancing historical insights with urgent calls to action. Consider a transgender scientist's reflections on his legacy or tag along on a quest to save a tiny porpoise from extinction. Crack open a history of immunology or confront the future of artificial intelligence. Why would 12 men dine on purposely poisoned foods? Can we overcome "chronophobia"? What can termites teach us about technology? Read on to discover these answers and more. —**Valerie Thompson**

The Beautiful Cure

By **Audrey R. Glynn**[1]

In *The Beautiful Cure*, immunologist Daniel Davis endeavors to tell the story of human immunology in eight chapters, spanning foundational concepts, pivotal discoveries by brilliant scientists and their collaborators, the personal drama that has accompanied Nobel prizes and other recognition awarded in this highly competitive field, and anecdotes about how individual choices can affect one's immune function and overall health.

Most readers will be familiar with the notion that stress increases vulnerability to illness. You may have noticed the way that pesky virus going around the office seems to skip over your meditation-practicing, well-rested co-worker, or perhaps you have experienced the reactivation of latent herpes virus (developed a cold sore) during demanding periods of your own life. In chapter 5, Davis skillfully tethers this phenomenon to the primary underlying factor: cortisol, a hormone that controls expression of >20% of human genes. Here, he describes how overproduction of cortisol during periods of stress effectively dampens innate immune responses, increasing the likelihood that exposure will result in infection rather than containment.

Cortisol is revisited in the subsequent chapter, which delves into Earth's cycles and circadian rhythms. Davis notes that in a healthy person, cortisol fluctuates predictably throughout the day and that, as a result, the performance of some drugs varies substantially depending on when they are administered. This phenomenon can be exploited for increased vaccine response. Davis estimates, for example, that "giving the [flu] vaccine in the morning would be able to protect over half of elderly people." One wonders, then, why drugs, by and large, are not prescribed for dosing at specific times of day, despite the fact that "fifty-six of the top hundred bestselling drugs in the USA ... target the product of genes that change their activity with the time of day."

In his attempt to tell a complete story, Davis makes a heroic effort to include all major discoveries and credit all the giants

ILLUSTRATION: ALEXANDER WELLS

upon whose shoulders the research community stands. In addition to the impact of stress and timing on immune function, he highlights major immune cells and mediators (such as dendritic cells and cytokines), delves into regulatory T cells and the hygiene hypothesis, and squeezes in a quick aside about the microbiome. However, given the complex, overlapping, and time-shifting nature of these discoveries (the first several pages of chapter 1 bounce the reader from 2008 to 1970 to 1721 to 1926), the narrative is often challenging to follow.

It is also unclear who Davis is writing for. Although he includes simplified descriptions of basic immunology concepts, seemingly to make the material accessible to a lay reader, he also describes complex scenarios without including any sketches or figures, from which such a reader would benefit greatly.

Overall, the book is of greatest value for biological scientists, for whom the relatively brief overview reveals intriguing connections in immunology's history, helps tie together stove-piped areas of inquiry, and offers fresh perspective on future research strategies.

The Beautiful Cure: The Revolution in Immunology and What It Means for Your Health, *Daniel M. Davis,* University of Chicago Press, 2018. 272 pp.

Vaquita

By **Todd L. Capson**[2]

In a poem for the recently extinct Yangtze River dolphin, *Goodbye Baiji*, author Brooke Bessesen laments the cetacean's untimely departure: "We bored of your life, your struggle—We tired of your incessant need, your slow demise. ... *What*? You are *gone*?... Now what will we do, not do?" The baiji's extinction in 2006 made Mexico's tiny endemic porpoise, the vaquita (*Phocoena sinus*), the "world's most endangered marine mammal." Unlike the Yangtze River dolphin, however, the vaquita may yet be saved. *Vaquita* is a lucid, informed, and gripping account of a species that will soon be lost in the absence of effective actions.

Although its populations are naturally small, with limited genetic diversity, lethal alleles have likely been purged, so the vaquita isn't biologically predisposed to extinction. Another sort of threat has rendered the species at risk.

Fishermen arrived at the Sea of Cortez in the 1920s in pursuit of *Totoaba macdonaldi*, a drum fish whose swim bladder provides a substitute for that of the critically endangered drum fish *Bahaba taipingensis*, which is coveted in China for its purported curative powers. A totoaba fisherman can earn up to $1800 for a single bladder, and in China, they can go for up to $250,000.

Since the 1940s, gillnets have been the tool of choice for capturing totoaba, but their unintended capture includes the vaquita, who often become entangled and drown. By the 1990s, as much as 90% of the original vaquita population was lost. From 1997 to 2008, the population plummeted another 57%. One Mexican scientist recently estimated that perhaps 15 remain.

In 1997, an advisory committee was formed, bringing together scientists and fisheries experts to provide input on the vaquita's conservation to the Mexican government. Bessesen describes the work of dedicated scientists from Mexico, the United States, and elsewhere and reveals that it is not for lack of data that the vaquita has suffered. She praises the work of Sea Shepherd, a nonprofit marine conservation group that assists in patrolling for poachers and safeguarding the Vaquita Refuge.

Vaquita details the tortuously slow and ineffective response of the Mexican government to the vaquita crisis. Past bans on gillnets, we learn, were often ignored or included loopholes that allowed their continued use. Plans to encourage fishermen to give up gillnets were ultimately unsuccess-

[1]The reviewer is at Strategic Analysis, Arlington, VA 22203, USA. Email: audrey.glynn@gmail.com [2]The reviewer is an independent consultant based in Dakar, Senegal. Email: capsont@gmail.com

A research trip to Svalbard impressed upon Marcia Bjornerud the amorphous nature of time.

ful, despite multimillion-dollar investments. Bessesen describes corruption at all levels of the government, exacerbated by widespread anti-vaquita sentiments.

The government of Mexico will ultimately determine the fate of the vaquita. The actions required to save the porpoise from extinction are clear, such as the need to make gillnets illegal to own and sell, a commitment to effective enforcement, and enabling the use of alternative fishing gears.

The Center for Biological Diversity has submitted petitions to the U.S. government to ban the import of Mexican seafood and other wildlife until the illegal totoaba trade ends. Despite China's commitment to cut off its consumers, very few arrests occur on that end of the supply chain. Perhaps the network behind the recent trade ban on ivory in China could next address totoaba bladders?

Bessesen approaches the plight of the vaquita with the thoroughness and inquisitiveness of a scientist and the passion of an environmentalist. She has written a must-read for anyone keen to understand the realities of protecting biodiversity. In doing so, she fulfilled a promise she made to a small female vaquita that died from entanglement in a gillnet: "I will tell your story."

Vaquita: Science, Politics, and Crime in the Sea of Cortez, *Brooke Bessesen,* Island Press, 2018. 313 pp.

Timefulness

By **David R. Wunsch**[3]

Earth, according to Marcia Bjornerud, is proportionally analogous to a peach. The pit is the core; the fruit is the mantle; the skin is the crust. And peach fuzz? It's the thickness of the atmosphere.

This is one example of the many wonderful analogies Bjornerud uses in *Timefulness* to help readers understand subjects ranging from Earth's structure to the concept of geologic time. Bjornerud's thesis is that having more people who can think like a geologist, rising above day-to-day concerns to conceive of the long-term consequences of our actions, will help society overcome many problems we face.

Bjornerud, an experienced field geologist, opens with a story about her first field research site on the Arctic archipelago of Svalbard, a location so remote and austere that it appears timeless. (Indeed, Svalbard literally has "No Official Time" because of a long-standing border disagreement.) It is from this benchmark that she gained an appreciation for the passing of time—a theme built on throughout the remainder of the book.

Bjornerud has a yeoman's command of historical geology, which she executes by moving through the evolution of our understanding of the age of Earth. She traces the postulations of icons of geology, including Charles Lyell and James Hutton, then moves on to the discovery and application of modern radiomentric dating and chronostatigraphy. She writes of the human aversion to the passage of time, which she terms "chronophobia," as epitomized by our narcissism and our obsession with defying our own ages.

Bjorneurd displays her wit in clever subchapter titles and headings. "Peak performances," for example, cheekily introduces a discussion of how erosion paradoxically causes mountains to rise in elevation. Quotes from ancient philosophers and notables throughout history add color and levity, while salient storytelling keeps the narrative informative without being tedious.

Black-and-white, hand-drawn sketches illustrate geologic concepts throughout the book. Although the figures complement the conversational writing style, they may be a bit hard to comprehend for those with less than a basic understanding of geology. One figure, for example, illustrates a distant view of a tidal flat featuring algal stromatolites. A block diagram with wavy lines represents a fossilized sample in the foreground. It may be difficult for nongeologists to ascertain how these two images interrelate.

Still, *Timefulness* is a delightful and interesting read. The author's cadence and the illustrator's aforementioned figures made me feel as though I was having a glass of wine with a friend who was explaining geologic history while sketching on a napkin.

Will *Timefulness* help save the world? Only time will tell.

Timefulness: How Thinking Like a Geologist Can Help Save the World, *Marcia Bjornerud,* Princeton University Press, 2018. 218 pp.

The Autobiography of a Transgender Scientist

By **David Litwack**[4]

In 2007, Stanford University neuroscientist Ben Barres published an essay comparing the experiences of female and male scientists. What made this essay noteworthy was

[3]*The reviewer is at the Delaware Geological Survey, University of Delaware, Newark, DE 19716, USA. Email: dwunsch@udel.edu* [4]*The reviewer is at Prevail Therapeutics, New York, NY 10011, USA. Email: edlitwack@gmail.com* [5]*The reviewer is at Scientific Planning Consulting, Highland Park, NJ 08904, USA. Email: daphnearobinson@gmail.com* [6]*The reviewer was at the Space and Advanced Technology Office, U.S. Department of State, Washington, DC 20520, USA. Email: ybutt2002@yahoo.com*

that he wrote from personal experience. Barres, an accomplished researcher and a tireless advocate for women in science, was assigned female at birth and transitioned to male in 1997 at the age of 43. Despite great personal and professional risks, he lived openly as a transgender scientist until his untimely death at the age of 63 from pancreatic cancer.

In his new, posthumously published book, *The Autobiography of a Transgender Scientist*, Barres breaks the account of his life into three sections: Life, Science, and Advocacy. His candor and love for science transform the ensuing story into a portrait of a singular personality that was shaped by his status as an outsider.

Barres grew up in a working-class family in New Jersey. He describes the dissonance of being raised as a girl, revealing the "continued emotional pain … that my gender discordance caused me." But during this period, Barres also developed a passion for science. After attaining a bachelor's degree from MIT, he went on to earn a medical degree and a doctorate in neurobiology from Dartmouth and Harvard, respectively.

Although Barres reports that these early years in his career were generally happy, he carefully catalogs the gender barriers and sexism he faced during this period. For example, upon solving a difficult math problem in a class at MIT, he recounts how the professor accused him of having had a boyfriend solve it. Such experiences informed his politics and drove him to champion the rights of female scientists.

Barres's passion for science does not seem to have been diminished by either the sexism he faced as a young academic or the transphobia he encountered later on. He routinely worked 16- to 20-hour days and once attempted a week-long vacation but left after 15 minutes on the beach, deciding he would rather be in the lab.

Barres devoted his scientific career to understanding the role of glial cells in the brain. This choice of subject is fitting: at the time he began his work, glial cells were also outsiders, widely perceived as unimportant. True to form, Barres persisted, and his findings played a big part in changing dogma. Today, glia are recognized as playing crucial roles in the wiring of the brain.

Some readers may find the science in this book inaccessible, which is unfortunate because it is a central element. A description of an experiment suggesting that "target innervation induced RGCs to down-regulate jagged1 mRNA," for example, sounds more like language used in a review article than in a memoir and fails to convey the drama of discovery or the full importance of his work. But for such lapses, Barres, who began writing this autobiography after his cancer diagnosis, should be forgiven.

In the end, although he reveals much that is insightful and important, the reader is left with the feeling that Barres had so much more to say.

The Autobiography of a Transgender Scientist, *Ben Barres,* MIT Press, 2018. 160 pp.

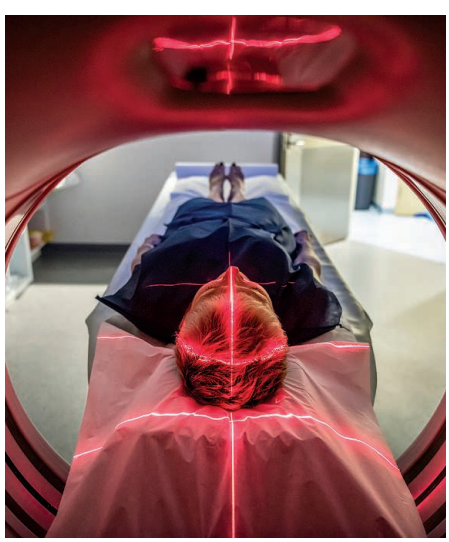
The New Mind Readers

By **Daphne A. Robinson**[5]

Mind reading is usually thought of as a magician's party trick. Yet advances in brain imaging have revived interest in this seemingly fictional feat. Can neuroimaging be used in court to show that a person is telling the truth or is in pain? What can neuroimaging tell us about what people think or how a specific person will behave?

In his book *The New Mind Readers*, Russell Poldrack addresses these and other tantalizing questions, presenting a clear and engaging overview of what neuroimaging can and cannot tell us about a person's thoughts, perceptions, and intentions. Going beyond basic mechanisms, Poldrack tackles a number of fundamental questions about the research process, data interpretation, and applications for everyday life.

Throughout the book, Poldrack presents the vast possibilities of neuroimaging while clearly articulating its limitations. Although it is possible, for example, to decode activity in the visual cortex in order to identify the general features of an image being viewed by an individual in a functional magnetic resonance imaging scanner, it is not possible to extrapolate an underlying emotion or mental state from brain activity if an individual is not performing a task specifically designed to elicit that emotion. Deducing a person's mental or emotional state solely on the basis of brain activity, a process called reverse inference, remains an important challenge that will require a more detailed understanding of how complex emotions are processed and represented throughout the brain and how brain activity gets combined across time and space.



Despite advances in neuroimaging, true mind reading remains, at present, more science fiction than fact.

PHOTO: ISTOCK.COM/JOHNNYGREIG

A fascinating set of issues emerge when we attempt to extrapolate research results to a single individual. Neuroimaging adds an additional layer of complexity: Each person's brain is a bit different, and brain activity can change over time. Poldrack outlines these limitations as he explores the implications of using neuroimaging data in a variety of settings, including the justice system, economic analyses, and marketing.

The reciprocal relationship between advances in neuroimaging and those in other fields is well illustrated in questions of addiction and mental illness. Addiction, as we now understand it, is both a disease that can be understood by imaging brain reward pathways and a highly context-dependent illness, indicating that activation of these pathways can be repressed. Similarly, although we now conceive of mental illness as a brain disease—a realization that has revolutionized our approach to mental illness—this has led some to wonder whether interventions can ever hope to override an individual's biology.

In the end, Poldrack is optimistic that the development of more specific imaging approaches should eventually enable at least a basic "dictionary" linking specific brain activity patterns to certain thoughts, activities, or emotions. Such insights, however, will likely be akin to the rudimentary comprehension of a newly arrived person in a foreign land, offering only a glimpse at a complex mental landscape.

The New Mind Readers: What Neuroimaging Can and Cannot Reveal About Our Thoughts, *Russell A. Poldrack,* Princeton University Press, 2018. 214 pp

Accessory to War

By **Yousaf Butt**[6]

In *Accessory to War*, Neil deGrasse Tyson and Avis Lang give a sweeping panoramic overview of the enduring alliance between astrophysics and the military—from the Greeks to Galileo to GPS. The book's key contribution is in documenting the various ways science has aided military endeavors over the millennia and making the sometimes-arcane source material accessible.

As the authors make clear, this isn't a one-way street, with science simply enabling greater military prowess or lethality. The military "Vela" satellites of the 1960s and 1970s, for example, were looking for γ-ray signatures from nuclear explosions on the ground during the Cold War. Instead they serendipitously found celestial γ-ray bursts coming from the other direction, ushering in γ-ray astrophysics.

And although space weapons may seem like a science fiction threat, they aren't; China, Russia, and the United States are all investing in space weaponization capabilities right now. Tyson and Lang do a good job of summarizing the state of play regarding the prospect of weapons in space, describing, for example, how the major powers point fingers at each other to justify acquiring ever greater capabilities. The authors outline the various diplomatic initiatives—draft treaties and codes of conduct—that have been put forth to keep space peaceful and why these efforts haven't met with much success so far.

In national security lingo, space is now cast as "congested, contested, and competitive," necessitating weapons to guarantee freedom of passage. But ever advancing science and technology create their own military imperatives in space. This echoes the assessment of Herb York, a prominent Cold War nuclear physicist and first chief scientist of the Advanced Research Projects Agency (ARPA), who, speaking about the arms race, memorably stated that, "the root of the problem has not been maliciousness, but rather a sort of technological exuberance that has overwhelmed the other factors that go into the making of overall national policy."

There's no question that national security imperatives have subsidized science, but they have also helped mold academic research. "Given the Cold War underpinnings of NASA's very existence, no astrophysicist should see NASA as our personal science-funding agency," we're told. "We are the wagging tail on a large geostrategic dog." Tyson and Lang sum it up succinctly: "Space exploration may pull in the talent, but war pays the bills."

The Cold War architecture seems to be functioning well enough, but what do the authors envision going forward? This is one of many topics discussed for which more critical analysis and the authors' personal views could have been enlightening. For those interested in a more analytic take on the interplay between science and the military, Daniel Sarewitz's "Saving science" is a good starting point (*1*).

REFERENCES

1. D. Sarewitz, *The New Atlantis*, spring/summer (2016); www.thenewatlantis.com/docLib/20160816_TNA49Sarewitz.pdf.

Accessory to War: The Unspoken Alliance Between Astrophysics and the Military, *Neil deGrasse Tyson and Avis Lang,* Norton, 2018. 590 pp.

Hello World

By **Arti Garg**[7]

In her opening explanation of the title to her book *Hello World*, mathematician Hannah Fry sets up a tantalizing question about the role of artificial intelligence (AI) in modern society by evoking the riddle about the chicken and the egg. In this instance, the riddle frames how we should think about the role that algorithms play in our lives. Who is in control? Humans? Or the machines we program to "think" like humans?

Fry explores these questions by discussing real-world applications of AI. Each chapter tackles a theme, such as "Crime" or "Art," in which she describes specific societal problems, reveals the algorithmic approaches being used to address them, and explains how those algorithms operate.

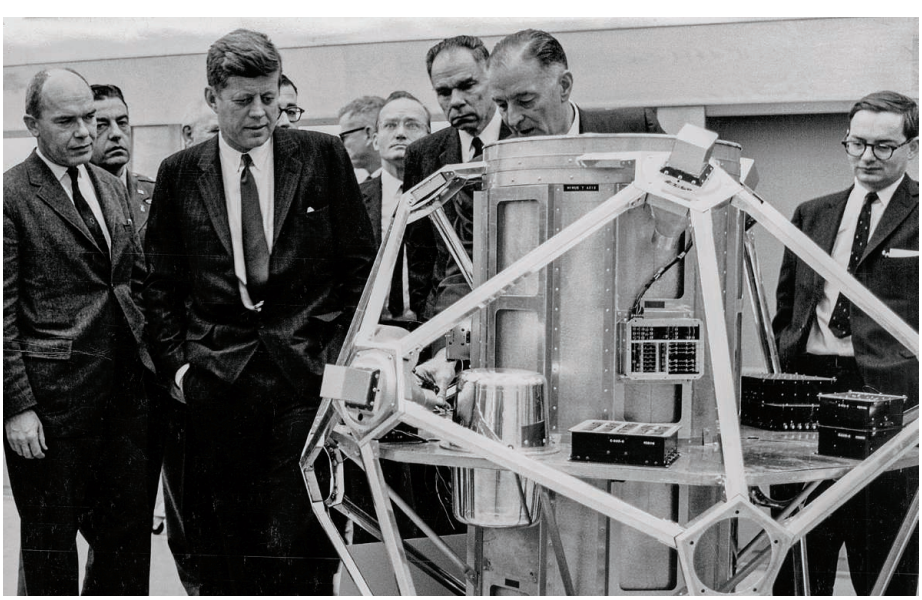
In this effort, Fry does yeoman's work by using concise and approachable prose to make the mathematics of AI accessible to a lay reader. Her analogy in the "Medicine" chapter, comparing a neural network to a machine with a series of tuning knobs, effectively conveys why it is not possible for us to interpret how these models work but doesn't oversimplify their underlying complexity.

After describing an algorithm and its application, Fry discusses its shortcomings. Unfortunately, in contrast to the vivid descriptions of the algorithms themselves, these discussions seem superficial, often leaving the reader without sufficient information to discern whether modifications to the algorithm could address the issues described or whether they are an unavoidable consequence of AI. This gap is most apparent in her section on "Justice."

Here, she raises the timely but controversial question of whether the use of AI in criminal justice proceedings perpetuates racial inequities. But instead of a concrete explanation of how to assess whether an algorithm is perpetuating biases and whether this can be corrected, she offers generalities about biased data. Citing expert opinion, the section culminates with Fry stating that despite its pitfalls, she considers the use of AI in the justice system preferable to its absence. This pronouncement feels rushed and unsatisfying, especially knowing that she is capable of more substantiated arguments.

Hello World concludes by imagining a world in which AI enhances human capabilities without overriding human judgment. Confusingly, this world is presented as an alternative to our current trajectory, even though—as Fry herself argues in the section "Cars"—it represents reality for the majority of today's AI applications. Such sweeping conclusions detract from the book's impact.

For a reader unfamiliar with the technical aspects of AI, this book offers among the best lay explanations of how algorithms work. But *Hello World* aspires to do more than this. It sets out to help us understand how to approach questions about the value and the unintended consequences of AI in our daily lives. Toward this end, the book, like the algorithms it describes, stops short of its promise. Despite its intriguing premise and



Vela satellites, designed to detect nuclear explosions, serendipitously ushered in γ-ray astrophysics.

PHOTO: SANDIA LABS/FLICKR/CC BY-NC-ND

broad-ranging subject matter, *Hello World* ultimately leaves the reader without a well-defined framework with which to evaluate the AI that they will encounter in the future.

Hello World: Being Human in the Age of Algorithms, *Hannah Fry,* Norton, 2018. 256 pp.

Underbug

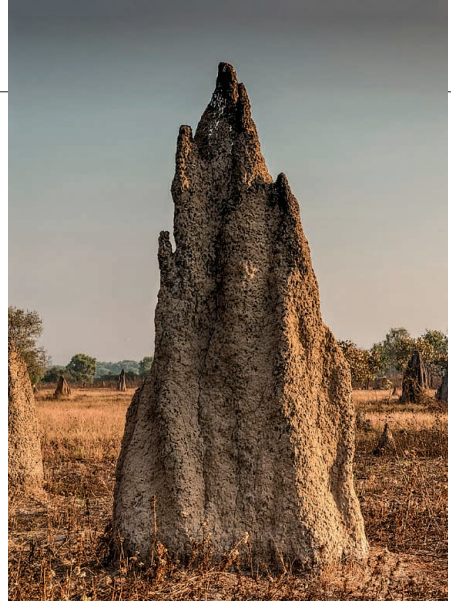
By **William J. Cromartie**[8]

Lisa Margonelli's *Underbug* book is definitely not about termites—at least, not as an entomologist would view them. Instead, it consists of stories of visits to labs and field sites, with reflections on questions raised by the research and researchers she encounters. Accounts of the biology of termites are scattered through chapters on how so-called "advanced" termites construct their complex mound nests, whether and how their collective behavior can be modeled and mimicked with artificial swarms of tiny robots, whether the metagenome of termite gut symbionts and their metabolic pathways might be engineered to make biofuels, and the role of termites in natural and restored ecosystems.

Margonelli sketches scientists at work and in moments of reflection, documenting their triumphs and failures. She contrasts nicely the different methods used by researchers to interrogate their subjects—for example, comparing how a team of molecular biologists tries to understand the significance of its data by playing a guessing game with the more systematic approach of a team led by a condensed matter physicist.

The book also has plenty to say about the nature of scientific inquiry and the strange ways that termites show humans in an unfamiliar perspective. To underline these points, Margonelli frequently references Eugène Marais's *The Soul of the White Ant* (1925), writing, for example, "His tale of the termite mound is part close observation, part poetic riddle, and part thumbnail guide to the universe, but it's not exactly scientific. Still, nobody since has gotten further into imagining the thoughts of the mound than Marais, making his book an invaluable document—of our minds more than theirs."

Comparing termites with artificial intelligence and miniature bioreactors, Margonelli describes familiar worries about the future of robotics and biological engineering, all while lamenting the loss of species and ecosystems. This is a lot to pile onto the backs of tiny insects. Are termite colonies really analogous to the human mind, as one interviewee—a physiologist studying mound construction—suggests?



Termites are not architects, writes Margonelli; their mounds are based on instinct, not a global vision.

Happily, termites stubbornly resist being reduced to mere robots or bioreactors. Far from being compulsively industrious, many appear to be slackers, to the surprise of the compulsively industrious scientists studying them. "You never get what you're looking for in biology," roboticist Kirstin Petersen laments in chapter 15. "We thought of every termite as the same termite…We were idiots."

The biggest weakness of *Underbug* is its structure: The book's various stories are told in roughly chronological chapters, so the narrative skips from topic to topic and place to place to such a degree that it becomes difficult to keep track of the various participants. Neither the termites nor the big issues hold the separate lines of inquiry together. That may make a point about science, but it leads to a rather confusing read. In the end, however, Margonelli has succeeded in presenting an interesting and provocative tale in which termites and people cross paths.

Underbug: An Obsessive Tale of Termites and Technology, *Lisa Margonelli,* Scientific American/Farrar, Straus and Giroux, 2018. 312 pp.

The Poison Squad

By **Melinda Cep**[9]

In 1844, when Harvey Washington Wiley was born, the federal agency at which he would eventually become chief chemist had not even been created. The U.S. Department of Agriculture was founded in 1862. Yet Wiley is still the hero of Deborah Blum's riveting, stomach-turning new book, *The Poison Squad*.

Whereas today as many as 15 federal agencies work with international, state, and local authorities to protect the American food supply, Wiley arrived on the scene when virtually no such protections existed. Before his retirement in March 1912, he and his team worked to implement some of the earliest food safety laws. But prior to doing so, they had to quantify how many of the food and drink products being consumed were adulterated (87% of the coffee samples tested, it turns out), determine the health effects of preservatives commonly found in food and drink (such as borax and salicylic acid), and communicate their findings to American consumers. Blum narrates the team's scientific and political adventures, including their era-specific difficulties (preserving food without modern refrigeration) and the more enduring challenges to consumer protection (obstructionism).

The Poison Squad offers a gripping history of the more than 20 years it took to pass the Pure Food and Drug Act of 1906, including all of the coalition building, advocacy, political failures, and cultural successes that accompany a major piece of legislation. Muckracking journalists such as Upton Sinclair aided Wiley's team's efforts, whereas backroom negotiations between companies and political figures stalled them. Budgets were slashed only to be restored; furious letters were drafted between Wiley and his various antagonists and allies; and an outpouring of support from American women and doctors ultimately helped Congress to pass the bill.

But passing a law isn't the same as getting it funded, implemented, and enforced. *The Poison Squad* offers an account of the complex relationship between a law, the appropriations to support its implementation, the rules to carry it out, regulatory decisions about enforcement, and subsequent legal challenges that may alter or undermine it all.

Blum isn't just telling one scientist's story but a broader one about the relationship between science and society. And because that relationship is maintained in much the same way today as it was in Wiley's time, *The Poison Squad* is a timely tale about how scientists and citizens can work together on meaningful consumer protections.

The Poison Squad: One Chemist's Single-Minded Crusade for Food Safety at the Turn of the Twentieth Century, *Deborah Blum,* Penguin Press, 2018. 352 pp.

10.1126/science.aav1600

[7]*The reviewer is at Cray, San Jose, CA 95112, USA. Email: arti.garg@stanfordalumni.org* [8]*The reviewer is at the Department of Environmental Studies, Stockton University, Galloway, NJ 08205, USA. Email: jamie.cromartie@stockton.edu* [9]*The reviewer is at the World Wildlife Fund, Washington, DC, 20037, USA. Email: melinda.cep@gmail.com*

PHOTO: SUMIO HARADA/MINDEN PICTURES/GETTY IMAGES

ECOLOGY

Learning to migrate

Hoofed animals, such as bighorn sheep and moose, learn migratory behaviors from other herd members

By **Marco Festa-Bianchet**[1,2]

Conditions for life are better in different places at different times of the year. Many animals—including birds, mammals, fishes, insects, and reptiles—take advantage of this temporal variation by migrating, sometimes over thousands of kilometers. But how do they know when and where to go, especially when released in an unfamiliar place? On page 1023 of this issue, Jesmer *et al.* (*1*) suggest that migratory ungulates—hoofed mammals—do something similar to that of tourists seeking local advice about places to eat: The ungulates' migration develops and persists through cultural transmission.

Spring starts earlier at lower altitudes and latitudes. A wave of green-up then moves higher or toward the poles, providing a gradient of highly nutritious vegetation over space and time. In mountains throughout the world, humans exploit this opportunity by moving livestock to higher grazing grounds from spring to summer. Reindeer herders also move their herds by hundreds of kilometers, imitating the long-distance migrations of tundra caribou. Wild migratory ungulates can feed over a long time on high-quality forage by following the onset of spring, or "surfing the green wave" (*2*). Better nutrition is one reason why migratory ungulates are much more numerous, and often physically larger, than sedentary ones, with important implications for ecology and the livelihoods of people that exploit them (*3*).

However, all migratory animals are extremely sensitive to human barriers to migration, such as fences and walls, including those at international borders. Because of the increasing prevalence of such barriers, many formerly migratory ungulates have become sedentary. In much of Europe, long-distance ungulate migrations likely disappeared centuries ago as humans changed the landscape. Migratory ungulates in Asia and Africa face ever more obstacles, and most bison in North America are inside large fenced areas because they are not tolerated on agricultural lands.

Because migratory ungulates are important ecologically and economically, governments, hunters, conservationists, and other people are interested in assisting their recovery, which requires knowledge on how migratory behavior develops. Jesmer *et al.* use a large sample of GPS-collared bighorn sheep and moose to investigate how the ability to migrate and surf the green wave varies with culturally transmitted local knowledge.

Most migratory ungulates are gregarious and likely learn when and where to go from their mother or other members of the herd. When a population is extirpated or reduced to so few animals that supplementation is required, translocation of animals from elsewhere may become a conservation option. But recently translocated ungulates must learn how to behave in their new environment because even short migrations may require crossing dangerous habitat. Bighorn sheep, for example, are at higher risk of predation away from the safety of cliffs, yet migration between seasonal ranges requires them to traverse risky habitats (see the photo and the figure). It turns out that to find the best seasonal ranges and surf the green wave, the animals mostly learn from others, sometimes over generations.

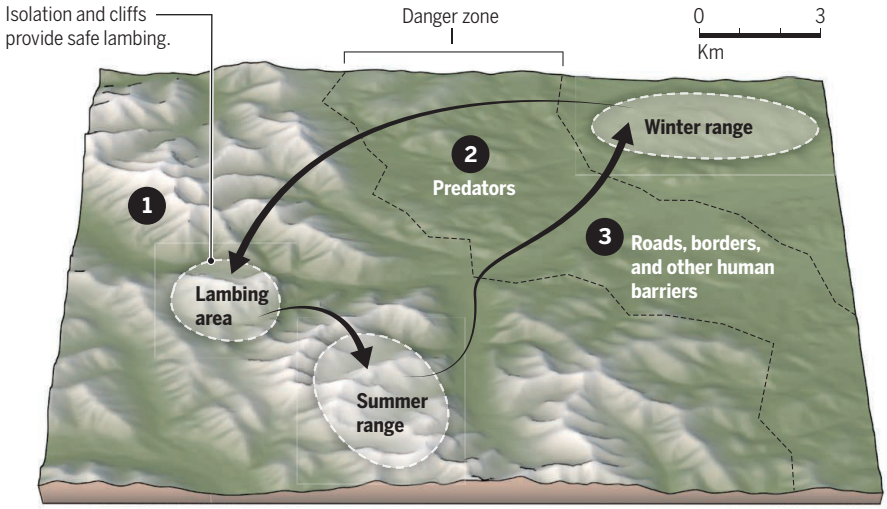
Migration requires knowing where the winter and summer ranges are. Surfing the green wave involves foraging at the best

[1]*Département de biologie, Université de Sherbrooke, Sherbrooke, Québec J1K 2R1, Canada.* [2]*Division of Evolution and Ecology, Research School of Biology, The Australian National University, Canberra, ACT 2601, Australia. Email: m.festa@usherbrooke.ca*

Bighorn sheep (*Ovis canadensis*) migrate in the mountains of the western United States.

Dilemma of a migratory bighorn ewe

Migratory bighorn sheep must find a safe route from the winter range to the lambing area and then the summer range. Jesmer *et al*. show that the sheep's migratory knowledge is culturally transmitted.



1 Seasonal changes
The location of the best forage sites changes with the seasons. It takes several generations to find the best forage sites, a skill called green-wave surfing.

2 Predation
During migration, the sheep must avoid predators such as wolves and cougars. The best routes follow the cliffs and avoid the forest.

3 Human barriers
Roads, fences, border walls, reservoirs, and railways can force detours, block migration, or increase mortality risk.

spots along the route each day of the migration. Jesmer *et al*. compared the migratory habits and green-wave surfing skills of bighorn sheep and moose from native populations, reestablished populations following translocations a few generations ago, and animals recently moved into new areas. Most animals in established populations migrated and generally had high green-wave surfing skills, selecting the best places at the best times. In recently established populations, migratory skills improved over time. Newly translocated animals did not migrate, except for a few that followed migratory residents. The authors show that it can take 90 years, or 12 to 13 generations, for half of the descendants of translocated animals to become migratory. Surfing skills were correlated with the development of migration but improved less markedly over time. Even in native herds, surfing performance was lower than that expected of a theoretical omniscient ungulate that surfed the green wave perfectly.

Translocations can be successful. The numbers and range of bighorn sheep in North America and ibex in Europe are much greater than a century ago, mostly through translocations. However, many translocations failed (*4*). Some failures were due to exotic disease or poaching, but others were likely associated with inability to reestablish historic migratory habits. Lack of local knowledge can also increase the risk of predation; avoiding predation is often another major advantage of ungulate migration (*3*). Jesmer *et al*. show that reestablishment of migratory habits is possible but takes time, especially in the absence of locally born conspecifics.

Migration and green-wave surfing do not correlate perfectly; some animals migrate between seasonal ranges but do not stop in all the best spots along the way. There are many reasons for imperfect green-wave surfing. One is predation. Animals may avoid areas with good forage but high predation risk (*5*). Another is the constraint imposed by the low mobility of newborns. Before giving birth, many ungulates move to sites that maximize safety over nutrition, migrating to foraging areas after juveniles have developed adequate locomotor skills. Last, whereas migratory ability simply requires a knowledge of seasonal range location, surfing the green wave must incorporate year-to-year changes in plant phenology, which may require repeated sampling of forage in different sites, with heightened predation risk and locomotory costs.

Jesmer *et al*.'s findings have important implications for conservation. After reintroductions, migratory ungulates may need several generations to learn the location of seasonal ranges. Learning from resident conspecifics speeds up the process but requires social integration, and translocated individuals may not be accepted quickly by the local population (*6*). Gregariousness likely plays an important role: The highly gregarious bighorn sheep may develop traditions more quickly than the mostly solitary moose. Future research may apply social network techniques to better understand who learns from who and how cultural transmission spreads through a population in which animals differ in levels of local knowledge. Given individual differences in migratory behavior and in ability to surf the green wave, some individuals may be more important than others in transmitting cultural knowledge about these fitness-enhancing behaviors. As populations dwindle, the chance that those individuals may be lost increases, possibly leading to loss of migratory behavior. Once local traditions are lost, it may take decades before they are reestablished.

Migratory ungulates, like other migratory animals, are important for many ecological processes and as a source of food, recreation, and cultural values for many people. Jesmer *et al*. underline the importance of culture in their ecology. That culture is lost when populations go extinct. Cultural transmission of migratory behavior is a major conservation challenge that can be best met through identification of critical migratory routes and habitat protection at a very large scale. ■

GRAPHIC: C. BICKEL/*SCIENCE*

REFERENCES

1. B. R. Jesmer *et al.*, *Science* **361**, 1023 (2018).
2. J. A. Merkle *et al.*, *Proc. R. Soc. B* **283**, 20160456 (2016).
3. J. M. Fryxell *et al.*, *Am. Nat.* **131**, 781 (1988).
4. F. J. Singer *et al.*, *Restoration Ecology* **8**, 6 (2000).
5. A. D. Middleton *et al.*, *Ecology* **94**, 1245 (2013).
6. M.-A. Poirier, M. Festa-Bianchet, *Biol. Conserv.* **218**, 1 (2018).

10.1126/science.aau6835

DEVELOPMENT

Matchmaking molecule for egg and sperm

Species-specific recognition molecule facilitates sperm binding to eggs at fertilization

By **Ruth Lehmann**

The union of egg and sperm occurs at fertilization; this creates a zygote that will develop into a new, unique organism carrying a mixture of the parental genomes. Thus, sexual reproduction, involving two separate gametes, diversifies the genetic pool and accelerates evolution. However, the added necessity for fertilization poses challenges. First and foremost, specific matchmaking is critical. Sperm must meet egg, and exactly one sperm needs to fertilize exactly one egg because an unfertilized egg cannot develop, and fertilization with too many sperm (polyspermy) is equally fatal. Furthermore, fertilization has to provide a barrier between species by creating egg-sperm incompatibilities. Despite the fascinating biological questions, evolutionary implications, and clear applications for human contraception and infertility inherent in understanding the mechanisms by which egg and sperm meet, these are still poorly understood. On page 1029 of this issue, Herberg *et al.* (*1*) investigate fertilization in zebrafish and identify Bouncer, a protein that is important for sperm-egg interaction and also creates a species-specific barrier. This discovery takes us a step closer to understanding species-specific fertilization and species evolution.

Fertilization is a highly orchestrated event (*2*). In preparation, both the egg and sperm must change: The sperm increases its mobility and the sperm head is reorganized to form a cap-like structure, called the acrosome. This prepares the sperm for penetration of the egg coat (the specialized extracellular matrix that surrounds the egg), attachment, and eventual fusion with the egg membrane. Egg activation leads to changes in the egg coat (also called the zona pellucida or chorion) (*3*). In mammals, the egg coat not only is a structural component that protects the oocyte but also attracts and binds sperm, whereas in fish, the micropyle, a specific opening in the egg coat, forms the single entry point for sperm. The egg coat becomes less permeable upon fertilization to prevent polyspermy. The fertilization process was first described in sea urchins by Oscar Hertwig in 1876 (*4*). Since then, many components of the sperm and egg machinery have been implicated in various aspects of the process. Surprisingly few of these have withstood the test of gene deletion experiments to demonstrate an essential function for egg-sperm recognition and fusion, defense against polyspermy, or species specificity (*2*).

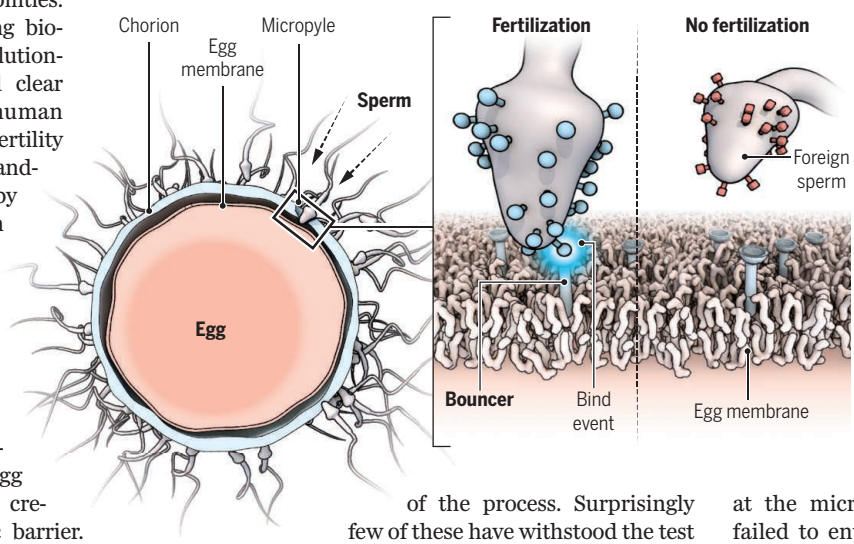
The discoveries of the sperm-specific factor Izumo (*5*) and its egg-specific receptor Juno (*6*) revealed the first pair of factors required for fertilization in mammals. Izumo, named after a Japanese shrine dedicated to the Shinto deity of marriage, encodes a single-pass transmembrane protein that binds via its N-terminal Izumo domain to Juno, named after the Greek goddess of marriage, which belongs to the folate receptor family and is anchored to the egg membrane via a GPI (glycosylphosphatidylinositol) linker (*7*, *8*). Beyond egg-sperm recognition, their exact role in fertilization has not been established. For example, although the pair mediates adhesion between sperm and egg or between heterologous cells, it is not sufficient to promote the fusion of two membranes in vitro (*6*). It has been proposed that vesicles expressing Juno on their membranes are extruded from the oocyte to function as an oocyte decoy to prevent polyspermy (*6*), possibly with the tetraspanin family protein CD9, but the mechanism is unknown (*9*). Izumo and Juno homologs exist throughout mammals, and their function in egg-sperm recognition could help discriminate against sperm from other species. However, evidence suggests that this receptor pair is not involved in the species barrier (*5*, *10*).

Searching for oocyte-specific expression of genes with unknown function in zebrafish (*11*), Herberg *et al.* discovered Bouncer, a short, GPI-linked membrane protein and member of the Ly6/urokinase plasminogen activator receptor family. Bouncer-deficient zebrafish are viable, but almost all females are completely sterile. The authors found no obvious defects in egg morphology, and egg activation seemed to proceed normally. Sperm accumulated at the micropyle, but the sperm nucleus failed to enter the egg when Bouncer was absent. The authors removed the egg coat and observed that sperm did not bind efficiently to Bouncer-deficient eggs.

To investigate whether Bouncer is involved in same-species recognition, Herberg *et al.* tested whether Bouncer can distinguish sperm from different species. They identified a homolog of Bouncer from a distantly related fish, medaka, which shares only 40% sequence identity with zebrafish-Bouncer. Unlike zebrafish-Bouncer, medaka-Bouncer did not restore fertility to Bouncer-deficient zebrafish when fertilized with zebrafish sperm. However, zebrafish eggs expressing medaka-Bouncer and fertilized with medaka sperm gained

Fertilization in fish

Bouncer is a receptor that is expressed on the outside surface of the egg membrane where it selects sperm from the same species, preventing fertilization by sperm from other species.



GRAPHIC: C. BICKEL/*SCIENCE*

Howard Hughes Medical Institute, Skirball Institute of Biomolecular Medicine, Department of Cell Biology, NYU School of Medicine, New York, NY, USA.
Email: ruth.lehmann@med.nyu.edu

some fertility (5.7% of eggs from fertile females produced embryos, compared with a fertility rate exceeding 78% when wild-type zebrafish were fertilized with zebrafish sperm). The resulting embryos were hybrids that carried a maternal zebrafish genome and a paternal medaka genome. The hybrid embryos survived early embryogenesis, underwent gastrulation, and formed an anteroposterior axis, but failed to develop past 48 hours. Thus, Bouncer not only is necessary for species-specific egg-sperm recognition, but is also sufficient to provide this function in the presence of sperm from another species (see the figure). The low fertilization rate when using medaka-Bouncer and medaka sperm is perhaps not surprising, given that all other components were zebrafish-specific and likely contributed to optimal recognition.

The closest homology to Bouncer in mammals is sperm acrosome membrane-associated protein 4 (SPACA4, also known as SAMP14), a GPI-linked protein that is expressed by sperm and thought to associate with the acrosome (*12*). Curiously, there seems to be sex-specific expression of Bouncer homologs according to the mode of fertilization (*1*). In fishes and amphibians, eggs are fertilized externally, so egg-sperm interactions are the first barrier for species recognition. In mammals, fertilization occurs internally with sperm traveling through the female reproductive tract. Interaction of sperm with somatic cells along the way to the oocyte has been thought to provide some species barrier (*2*). Perhaps sperm-expressed Bouncer is a species-specific cue that is recognized by somatic cells prior to egg encounter. Clearly, characterization of phenotypes associated with SPACA4 deficiency will provide new insight. Additionally, discovering the counterpart or counterparts that are bound by Bouncer on the sperm of fishes, and on the eggs and possibly somatic cells lining reproductive tracts in mammals, will help us to understand this puzzle. ■

REFERENCES

1. S. Herberg *et al.*, *Science* **361**, 1029 (2018).
2. E. Bianchi, G. J. Wright, *Annu. Rev. Genet.* **50**, 93 (2016).
3. P. M. Wassarman, E. S. Litscher, *Curr. Top. Dev. Biol.* **130**, 331 (2018).
4. O. Hertwig, *Morphol. Jahrb.* **1**, 347 (1876).
5. N. Inoue *et al.*, *Nature* **434**, 234 (2005).
6. E. Bianchi *et al.*, *Nature* **508**, 483 (2014).
7. H. Aydin *et al.*, *Nature* **534**, 562 (2016).
8. U. Ohto *et al.*, *Nature* **534**, 566 (2016).
9. M. Chalbi *et al.*, *Development* **141**, 3732 (2014).
10. E. Bianchi, G. J. Wright, *Philos. Trans. R. Soc. B* **370**, 20140101 (2015).
11. A. Pauli *et al.*, *Science* **343**, 1248636 (2014).
12. J. Shetty *et al.*, *J. Biol. Chem.* **278**, 30506 (2003).

10.1126/science.aau8356

NEURODEGENERATION

A brain boost to fight Alzheimer's disease

A mouse model of Alzheimer's disease provides clues about why exercise is good for memory

By **Tara L. Spires-Jones**[1] *and* **Craig W. Ritchie**[2]

Alzheimer's disease is one of the biggest medical challenges of our time. About 30 million people worldwide are living with Alzheimer's disease, and the numbers are predicted to increase to 100 million by 2050 if we do not find effective prevention or treatment strategies (*1*). Substantial evidence suggests that leading a healthy lifestyle, including regular exercise, may lower the risk of developing Alzheimer's disease. However, the mechanisms through which exercise protects the brain and whether we could bottle these as a treatment remain controversial. On page 991 of this issue, Choi *et al.* (*2*) reveal that in a mouse model of Alzheimer's disease, exercise improves memory through a combination of encouraging neurogenesis in the hippocampus and increasing the levels of brain-derived neurotrophic factor (BDNF), a growth factor that supports neuronal survival. Their findings suggest that agents that promote both BDNF signaling and neurogenesis might be effective in preventing or treating Alzheimer's disease.

> ***"...the beneficial effects of exercise...are due to enhancing neurogenesis in combination with increasing BDNF amounts in the brain."***

People with Alzheimer's disease have accumulations of pathological lesions called amyloid plaques and neurofibrillary tangles in their brains, along with dramatic loss of synapses and neurons, that cause severe cognitive decline (*3*). Epidemiological studies show that exercise reduces the risk of cognitive decline during aging and decreases the risk of Alzheimer's dementia (*4, 5*). However, the evidence from clinical trials of exercise in people has been inconsistent, tending to no net benefit on symptoms in people with dementia (*6*) and no benefit on cognition as a preventive measure in healthy elderly individuals (*7*). This suggests that there is a complex relationship between exercise and brain health that depends on life stage. The mechanisms leading from exercise to protecting the brain are not well understood, which undermines the optimal design of interventional studies as well as public health advice and what the exercise intervention should be—for example, duration, frequency, aerobic versus anaerobic, and when this intervention should be delivered for maximal benefit.

Several hypotheses have been proposed to explain the preventive effects of exercise, including the idea that exercise boosts neurogenesis, which consequently enhances brain function. To explore this idea, Choi *et al.* used a mouse model of Alzheimer's disease that overexpresses gene mutations associated with rare familial forms of the disease and develops amyloid pathology and memory deficits. Choi *et al.* found that the beneficial effects of exercise—in this case, voluntary aerobic exercise on a running wheel for 3 hours per day—are due to enhancing neurogenesis in combination with increasing BDNF amounts in the brain (see the figure).

This work is important because it provides preclinical proof of concept that a combination therapy that increases neurogenesis and BDNF levels could be disease-modifying or prevent Alzheimer's disease altogether. Importantly, in these mice, memory could be improved by using a combination of molecular interventions without the need for exercise. Neurogenesis was induced with two factors: a viral injection in the hippocampus to express the protein WNT3 combined with systemic treatment with P7C3, a compound that improves survival of new neurons. BDNF levels were raised either by injecting a virus expressing the protein into the hippocampus or by pharmacological means, with systemic injection of 5-aminoimidazole-4-carboxamide riboside (AICAR). Although the specific viral interventions used are unlikely to be practical for preventing Alzheimer's disease in people, the idea that pharmacological interventions

[1]*UK Dementia Research Institute, Centre for Discovery Brain Sciences, The University of Edinburgh, Edinburgh, UK.* [2]*Centre for Clinical Brain Sciences, The University of Edinburgh, Edinburgh, UK. Email: tara.spires-jones@ed.ac.uk*

can mimic some of the effects of exercise is promising. This is particularly important for older people who might not always have the capacity for the level of exercise needed to promote optimal brain health.

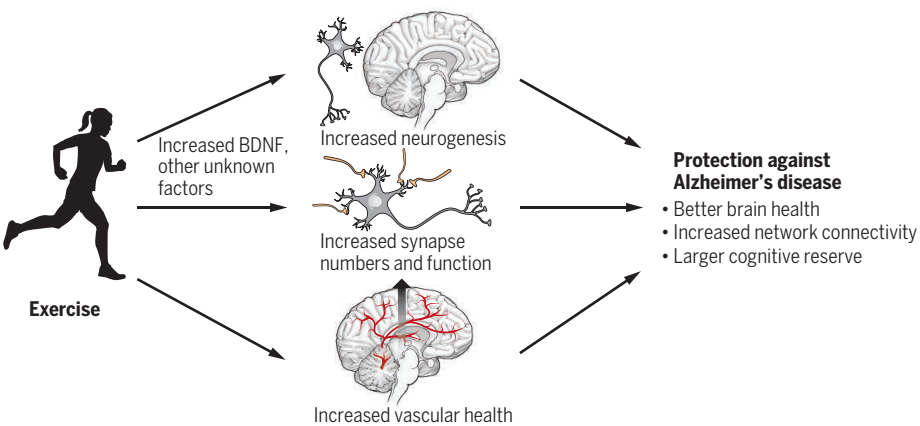
The study by Choi *et al.* also lends further support to the idea that amyloid plaques are a poor marker of disease progression and thus a poor choice of biomarker for clinical trials. The authors found that decreased plaque pathology was not necessary or sufficient to drive memory improvements. This adds to conflicting reports that environmental enrichment, including exercise, in some studies decreased plaque pathology and increased it in others, despite most groups observing memory improvement (*8, 9*). Biomarkers that more closely track neural circuit integrity are much more promising, including structural magnetic resonance imaging to track brain atrophy and newer markers such as positron emission tomography ligands for synapses and synapse and neuron proteins in cerebrospinal fluid (*10*).

There are limitations of this work to consider. The authors used a single aggressive mouse model of familial Alzheimer's disease, which potentially exhibits nonphysiological mechanisms of cell death owing to overexpression of mutant proteins. Historically, studies of mouse models of Alzheimer's disease have not translated well in human clinical trials, particularly work in a single model, so this study will need to be replicated in other mouse models. There is also no direct evidence that these mechanisms are involved in human disease. Neurogenesis in human adult hippocampus has become a contentious issue. For example, one study found that neurogenesis occurred into older age (*11*), whereas another found that it stops during childhood (*12*). Overall, evidence suggests that at least low levels of neurogenesis occur throughout life. There have been several studies indicating that increased neurogenesis occurs in Alzheimer's disease (*13*), but whether this plays an important role in disease pathophysiology is unclear.

Moving forward, it will be important to understand how exercise, neurogenesis, and BDNF affect the brain at the synapse, cellular (neurons, glia, and vascular cells), and circuit level. Exercise causes new synapse formation and is excellent for cardiovascular health, both of which are relevant to Alzheimer's disease because synapse degeneration is an important correlate of cognitive decline, and we are beginning to understand that nonneuronal cells (glia and vasculature) greatly affect this process (*3, 14*). In the best-case scenario, assuming these results are replicated in other models and are relevant to human disease, this study suggests that we could bottle the effects of exercise to prevent or treat dementia. ■

How might exercise protect against Alzheimer's disease?

Several pathways might explain how exercise protects the brain and prevents development of Alzheimer's disease. In mice, exercise enhances vascular health and increases the amount of BDNF in the brain, which promotes neurogenesis, survival of new neurons, and the formation of new synaptic connections.



REFERENCES

1. WHO, "Dementia: a public health priority" (World Health Organization, 2012); www.alz.co.uk/WHO-dementia-report.
2. S. H. Choi, *Science* **361**, eaan8821 (2018).
3. T. L. Spires-Jones, B. T. Hyman, *Neuron* **82**, 756 (2014).
4. M. Hamer, G. Muniz Terrera, P. Demakakos, *J. Epidemiol. Community Health* **72**, 477 (2018).
5. G. Livingston *et al.*, *Lancet* **390**, 2673 (2017).
6. D. Forbes, S. C. Forbes, C. M. Blake, E. J. Thiessen, S. Forbes, *Cochrane Database Syst. Rev.*, CD006489 (2015).
7. J. Young, M. Angevaren, J. Rusted, N. Tabet, *Cochrane Database Syst. Rev.*, CD005381 (2015).
8. O. Lazarov *et al.*, *Cell* **120**, 701 (2005).
9. J. L. Jankowsky *et al.*, *J. Neurosci.* **25**, 5217 (2005).
10. M.-K. Chen *et al.*, *JAMA Neurol.* 10.1001/jamaneurol.2018.1836 (2018).
11. M. Boldrini *et al.*, *Cell Stem Cell* **22**, 589 (2018).
12. S. F. Sorrells *et al.*, *Nature* **555**, 377 (2018).
13. D. Gomez-Nicola *et al.*, *Brain* **137**, 2312 (2014).
14. B. De Strooper, E. Karran, *Cell* **164**, 603 (2016).

10.1126/science.aau8060

PLANT BIOLOGY

To grow and to defend

A rice transcription factor's dual roles could improve yield and disease resistance

By **George H. Greene** *and* **Xinnian Dong**

Feeding an expanding world population while sustaining an inhabitable environment represents the greatest challenge of our time. To meet this challenge, the scientific conundrum of increasing crop yield while protecting it from evolving pathogens must be resolved. Rice (*Oryza sativa*) contributes the majority of dietary energy for more than half of the world's population. The most devastating pathogen of rice worldwide is the fungus *Magnaporthe oryzae*, the causal agent of rice blast, which results in an estimated yield loss of 30% globally. Therefore, controlling *M. oryzae* infection is a key battlefront for improving global rice production (*1*). On page 1026 of this issue, Wang *et al.* (*2*) identify a mechanism by which Ideal Plant Architecture 1 (IPA1), a transcription factor previously identified for conferring high yield, can also promote immunity against rice blast. This discovery provides a great addition to the toolbox for rational breeding of rice varieties with both high yield and high disease resistance.

Disease resistance is mostly generated by introducing immune receptor genes from resistant cultivars or wild relatives into high-yield varieties to create "supervarieties" (*3*). Many of these immune receptors recognize the presence of rapidly evolving, pathogen-specific virulence factors. Selection for new polymorphisms that allow pathogen effectors to evade perception can diminish the efficacy of these receptors within years (*4*). A common remedy for this deficiency is the stacking of multiple immune receptor genes in the host genome (*5*). However, expression of these receptor genes can lead to competition with yield-related traits for available metabolic resources (*6*). Alternatively, broad-spectrum disease resistance may be conferred through manipulation of central immune regulators downstream of pathogen recognition that control the expression of resistance-conferring genes. Constitutive

Howard Hughes Medical Institute, Department of Biology, Duke University, Durham, NC 27708, USA. Email: xdong@duke.edu

GRAPHIC: N. DESAI/*SCIENCE*

activity of these core regulators is often associated with drastically diminished yield because of the inherent trade-offs between defense response and growth (*7*). Recent advances in engineering broad-spectrum disease resistance have focused on mitigating yield costs by making the immune response tunable to pathogen detection. Such strategies include the use of pathogen-responsive promoters (*8*), messenger RNA (mRNA) stability (*9*), and chromatin modifications (*3*). More recently, regulation at the level of translation was achieved through cis mRNA elements that allow transient translation of core immune regulators to establish optimal balance between defense and growth (*10*).

Although considerable effort has gone into understanding and minimizing the effects of immune regulators on growth, the extent to which growth-promoting regulators modulate immunity remains relatively unexplored. Yield is a complex trait that is determined by multiple factors, including the number of panicles (branched flower clusters), the number of filled grains per panicle, and overall grain weight (*11*). IPA1 functions to reduce the number of unproductive rice tillers (panicle-bearing stems) and increase grain density in productive panicles (*12*), maximizing the number of grains per panicle and thus the yield of the plant. Alleles of *IPA1* from various cultivars are expressed at different amounts owing to variations in distinct regulatory mechanisms, including methylation of the *IPA1* promoter in the *wealthy farmer's panicle* (*WFP*) allele or microRNA (miRNA)–mediated repression of *IPA1* expression. The *ipa1-1D* allele contains a naturally occurring point mutation that escapes the miRNA-mediated repression, resulting in higher IPA1 protein expression (*13*). Fine-tuning IPA1 protein abundance allows for the ideal combination of panicle size and panicle number for maximum productivity. However, the value of such high-yield varieties is limited if losses to disease are not controlled.

Wang *et al*. investigated the yield output of *ipa1-1D* plants under *M. oryzae* challenge. Through large-scale field trials, they determined that plants with the *ipa1-1D* allele maintained the expected yield benefits of ~10% (compared with plants expressing the *IPA1* allele) in the noninfected field and showed an impressive 30% yield increase in the infected field. This led the authors to hypothesize that IPA1 promotes resistance to rice blast in addition to its yield-enhancing activity.

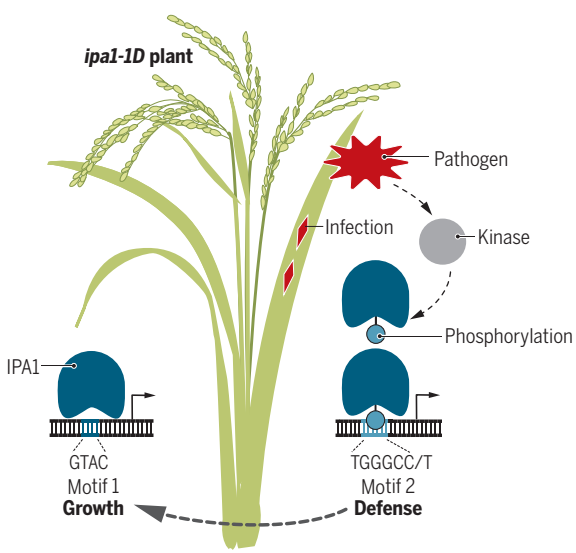
Through biochemical assays, the authors identified an inducible phosphorylation event on serine-163 in a conserved region of IPA1 that occurs upon exposure to *M. oryzae*. In the absence of pathogen, IPA1 binds to the promoters of growth-stimulating genes to drive their expression. Phosphorylated IPA1 is repurposed as its DNA binding affinity is altered to favor defense gene promoters, including the promoter of WRKY DNA-binding protein 45 (WRKY45), which encodes a broadly pathogen-responsive transcription factor involved in global transcriptional changes associated with plant defense (*14*, *15*). The increased abundance of IPA1 in the *ipa1-1D* plants accelerates WRKY45 accumulation, conferring enhanced resistance.

Moreover, phosphorylated IPA1 in *ipa1-1D* plants returned to background-level amounts by ~48 hours after infection. This could result from turnover of the phosphorylated IPA1 protein or an active dephosphorylation process when infection subsided. This rapid and transient IPA1 modification allowed *ipa1-1D* plants to have increased yield and enhanced *M. oryzae* resistance while avoiding the yield penalties that would occur if IPA1 remained in its immune active form (see the figure). It warrants further investigation whether such pathogen-responsive phosphorylation of growth-regulating transcription factors is a widespread mechanism for rapid yet moderated response to infection.

The dual functionality of IPA1 presents intriguing new opportunities for engineering disease resistance to other pathogens and in additional crops. This depends on the degree of conservation of the signaling components, from pathogen recognition to IPA1 phosphorylation to the downstream defense targets. For example, if IPA1 phosphorylation is dependent on pathogen recognition through a specific immune receptor, the use of IPA1 may be limited to *M. oryzae* in rice. Nevertheless, further engineering through introduction of an IPA1 overexpression allele, along with the necessary upstream components, may confer yield benefits that are retained under challenge by a broad spectrum of pathogens. ■

How IPA1 promotes growth and resistance

IPA1 normally binds GTAC DNA sequences (motif 1) to promote growth. Upon *M. oryzae* infection, IPA1 is phosphorylated by an unknown kinase and switches its binding to TGGGCC/T (motif 2), which promotes immunity to *M. oryzae*.



After immune induction, IPA1 rapidly reverts to its growth-promoting function (gray arrow), mitigating potential yield losses associated with unregulated immune responses.

REFERENCES

1. S. Ashkani *et al*., *Front. Plant Sci.* **6**, 1 (2015).
2. J. Wang *et al*., *Science* **361**, 1026 (2018).
3. Y. Deng *et al*., *Science* **355**, 962 (2017).
4. R. Panstruga, P. N. Dodds, *Science* **324**, 748 (2009).
5. S. Fukuoka *et al*., *Sci. Rep.* **5**, 1 (2015).
6. J. Bergelson, C. B. Purrington, *Am. Nat.* **148**, 536 (1996).
7. H. Fitzgerald *et al*., *Mol. Plant Microbe Interact.* **2**, 140 (2004).
8. J. Vijayan *et al*., *Front. Plant Sci.* **6**, 371 (2015).
9. X. Zhou *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **115**, 3174 (2018).
10. G. Xu *et al*., *Nature* **545**, 491 (2017).
11. D. Zeng *et al*., *Nature* **3**, 4 (2017).
12. L. Zhang *et al*., *Nat. Commun.* **8**, 14789 (2017).
13. B. Wang, H. Wang, *Mol. Plant*. **10**, 779 (2017).
14. M. Shimono *et al*., *Plant Cell* **19**, 2064 (2007).
15. Y. Ke *et al*., *Plant J.* **90**, 738 (2017).

10.1126/science.aau9065

COMBUSTION CHEMISTRY

A radical approach to soot formation

Chain reactions of resonance-stabilized radicals turn molecules into clusters and nanoparticles

By **Murray Thomson** *and* **Tirthankar Mitra**

Humans have used the high-temperature synthesis of carbon particles as a source of pigments since prehistoric times (*1*), and today, "carbon blacks" are used in tires, inks, coatings, plastics, and electrical applications. For most substances, solid particles become gases when heated, but solid soot forms from gaseous molecules at high temperature through a mechanism that is still not understood. The high-temperature synthesis of carbon particles (see the figure) occurs under oxygen-starved conditions where simple hydrocarbons can grow into larger molecules, especially polycyclic aromatic hydrocarbons (PAHs), in the gas phase. These large molecules cluster into carbon nanoparticles, a process often referred to as soot inception. The carbon nanoparticles then grow and aggregate to form larger, fractal-like structures. On page 997 of this issue, Johansson *et al.* (*2*) propose a new mechanism based on chain reactions of resonance-stabilized radical (RSR) species to explain this transition from gas-phase molecules to nanoparticles.

Most models of soot inception use the approach of Frenklach and Wang (*3*), who modeled soot nucleation based on the irreversible collision of PAH species that link together to form a dimer. A dimer would be better described as a molecular cluster rather than soot, so it is not possible to define a clear moment when the growing cluster becomes soot. Thus, many models group dimers, clusters, incipient soot, and mature soot into the soot model. In later work, Eaves *et al.* (*4*) modeled the formation of the dimer as a reversible process to be consistent with thermodynamics. However, many researchers have noted that PAH dimers held together by physical forces (e.g., van der Waals interactions) are not stable at high temperatures (*5*), and the PAH dimers will break apart.

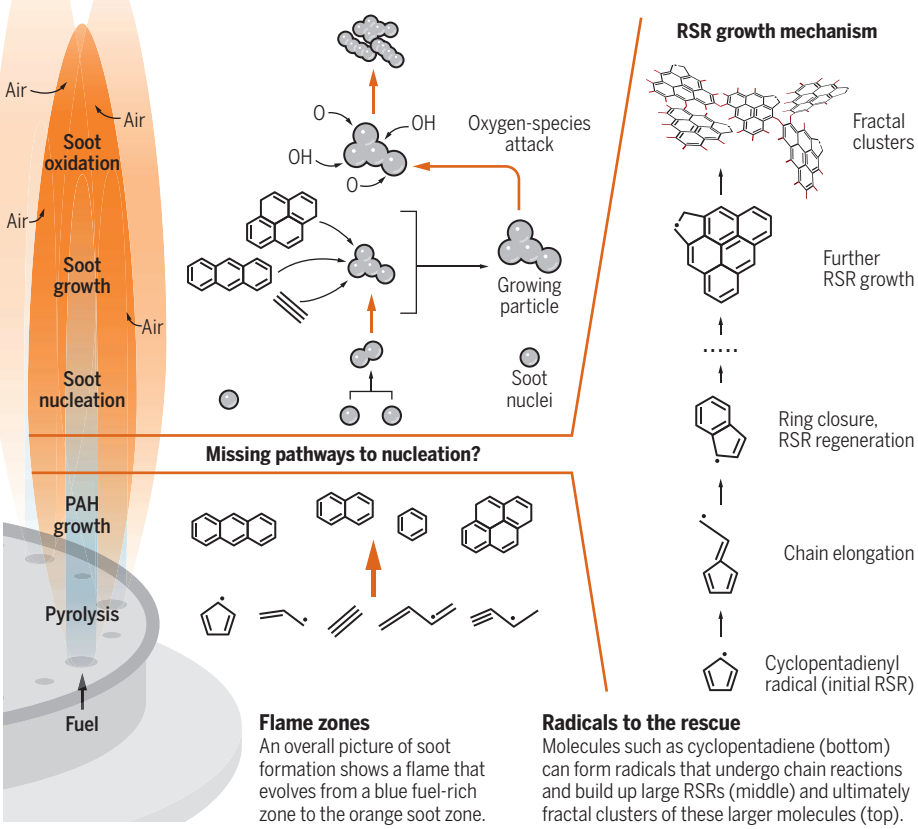
Thus, chemical bonds between the PAH molecules in the cluster or particle are essential to maintain particle stability at the high flame temperatures. Several researchers have proposed pathways for covalent bond formation, but these pathways use irreversible reactions with high-energy barriers. The challenge is to provide a detailed reaction mechanism that explains the rapid clustering yet is consistent with fundamental thermodynamics principles.

Johansson *et al.* propose a new mechanism to overcome this challenge and provide pathways to rapid particle formation that are fully consistent with fundamental scientific principles. The CHRCR (clustering of hydrocarbons by radical-chain reactions) mechanism invokes chain reactions of RSR species. These RSR species are radicals and highly reactive. Every step in this sequence forms a more stable RSR, and thus the reaction moves forward. The chain reactions can begin with small RSR species such as the cyclopentadienyl radical. The species growth is fast because the reactions of these RSRs with other radical and closed-shell hydrocarbons readily regenerate larger RSRs. There is no need for additional hydrogen-abstraction reactions that can have a high energy barrier. The large resonantly stabilized π-radical can then undergo σ-dimerization to form a molecular cluster.

Although this dimerization process forms a stable cluster, the cluster can regain its radical character by losing an H atom and form products with higher stability. Thus, the radical pool never gets depleted. The proposed reaction steps are very exothermic (they release energy) and should be effectively irreversible. The cluster can continue to grow via reactions with other RSRs, unsaturated aliphatic species, or nonradical aromatic species. This process would result in the fast and irreversible growth of large molecular clusters that eventually leads to soot particles.

From gas to soot

The initial formation of solid nanoparticles from gas molecules is a missing pathway in soot formation. Johansson *et al.* propose a mechanism in which chain reactions of resonance-stabilized radicals (RSRs) account for the growth of polyaromatic hydrocarbons (large RSRs) and their cluster growth.



Flame zones
An overall picture of soot formation shows a flame that evolves from a blue fuel-rich zone to the orange soot zone.

Radicals to the rescue
Molecules such as cyclopentadiene (bottom) can form radicals that undergo chain reactions and build up large RSRs (middle) and ultimately fractal clusters of these larger molecules (top).

GRAPHIC: A. KITTERMAN/*SCIENCE*

Department of Mechanical and Industrial Engineering, University of Toronto, Canada. Email: murray.thomson@utoronto.ca

The CHRCR mechanism proposed by Johansson *et al.* may eventually resolve the mystery of how soot inception occurs. However, to test this hypothesis, there is a need to establish detailed reaction rates for the many reactions between RSRs and other radical and stable hydrocarbon species, as well as for the reactions forming the initial clusters and the growth pathways of these clusters. Further research is also necessary to identify how these clusters transition into three-dimensional solid structures. Once these reaction rates are determined, the model predictions can be validated quantitatively against soot and gas-phase species measurements in laboratory reactors and flames. These laboratory experiments will need to span the wide range of gas compositions, temperatures, and pressures in which these carbon nanoparticles are formed in practical applications. This process is likely to iterate between model improvement and validation. However, the short lifetime of the RSRs may make their experimental measurement more challenging.

The issue of carbon particle formation is not only of industrial interest. Soot or smoke is unintentionally formed in combustion applications and in wildfires, and if released to the environment, soot contributes to air pollution and climate change. Space dust in interstellar regions is believed to form from carbon-containing stars through a mechanism similar to that for soot (*6*). Now that Johansson *et al.* have proposed an innovative and promising new approach to understanding soot inception, the hard work will be filling in the details of the model that should help solve a mystery that has challenged researchers for at least the last 40 years. ■

"Further research is also necessary to identify how these clusters transition into three-dimensional solid structures."

REFERENCES

1. M.-J. Wang, C. A. Gray, S. R. Reznek, M. Mahmud, Y. Kutsovsky, *Carbon Black*, in *Encyclopedia of Polymer Science and Technology* (Wiley, 2003); https://doi.org/10.1002/0471440264.pst477.
2. K. O. Johansson, M. P. Head-Gordon, P. E. Schrader, K. R. Wilson, H. A. Michelsen, *Science* **361**, 997 (2018).
3. M. Frenklach, H. Wang, *Symp. Int. Combust.* **23**, 1559 (1991).
4. N. A. Eaves, S. B. Dworkin, M. J. Thomson, *Proc. Combust. Inst.* **35**, 1787 (2015).
5. T. S. Totton, A. J. Misquitta, M. Kraft, *Phys. Chem. Chem. Phys.* **14**, 4081 (2012).
6. A. G. G. M. Tielens, *Annu. Rev. Astron. Astrophys.* **46**, 289 (2008).

10.1126/science.aau5941

ASTROPHYSICS

Computational astrophysics for the future

An open, modular approach with agreed standards would facilitate astrophysical discovery

By **Simon Portegies Zwart**

Scientific discovery is mediated by ideas that, after being formulated in hypotheses, can be tested, validated, and quantified before they eventually lead to accepted concepts. Computer-mediated discovery in astrophysics is no exception, but antiquated code that is only intelligible to scientists who were involved in writing it is holding up scientific discovery in the field. A bold initiative is needed to modernize astrophysics code and make it transparent and useful beyond a small group of scientists.

Scientific software is like a prototype in a laboratory experiment; it must stimulate experimentation. The eventual code is a description of concepts and their relationships, which are imperative for reproducibility and validating the results. In recent decades, hardware performance and computer languages have both changed dramatically. Software engineering has also experienced major advances, such as multilingual implementations and the introduction of design patterns. Industry and large scientific endeavors, such as CERN, have embraced these changes; in addition, their software is built on various assisting packages to vitalize modularity, platform independency, and message brokering. Although leading to a better structure, these developments also lead to higher complexity (*1*) and a sharp increase in the number of code lines (*2*). The latter is particularly noticeable in the growth of the Linux kernel (*3*).

Astronomical source code remains tiny by industrial standards, and its structure is characterized by developments during the "software crisis" of 1965 to 1985, when software was written as a long list of instructions without formal structure (*4*). The relative simplicity of these "dinosource" codes facilitates their survival but frustrates further development. Scientific source code is experimental in much the same way as laboratory experiments; it is not a final concept and is never ideally organized because it is intended to mediate exploration. By contrast, industrial code is mature but restrains experimentation. Furthermore, industry can afford dedicated teams to design and maintain software, whereas astronomical software development is organized in indigent "families" of researchers. A modular approach with agreed standards is essential to embolden astrophysical discovery by computer.

Leiden Observatory, Leiden University, 2300 RA Leiden, Netherlands. Email: spz@strw.leidenuniv.nl

CITING AND SHARING

The simple structure of astronomical software packages has enabled them to survive, some even since the 1970s. Although this dinosource code is often written in an ancient dialect, the underlying physics has hardly changed. About 58% of these codes are publicly available (*5*), but many are obfuscated because they have received multiple updates in the form of patches, keeping the deprecated code for backward compatibility. Existing code therefore forms no suitable basis for building a general framework. We know little about undisclosed codes, but it is unlikely that they are different.

Authors may decide not to distribute their source code; this is generally motivated by a desire to preserve a head start on the competition, or for fear that erroneous results produced by others discredit their efforts. Such secretly developed codes are of no help to the community and produce unverifiable results. In experimental sciences, failure to publish laboratory details is equivalent to failure in properly describing an experiment (*6*), and this cannot be much different for source code. But so long as public codes are copied, adapted, or expanded and eventually released under a different name, one can hardly expect the community to change its behavior. The apparent lack of trust that sufficient credit is given to code development can only be lifted by enabling codes to be cited directly.

The Astrophysics Source Code Library (*7*) provides a platform that mediates sharing and citing of astronomical source code, but it does not assign a digital object identifier (DOI). Such a service is provided by Zenodo (*8*), and all ingredients to organize the community are in principle available.

A specialized journal that assigns a DOI to a code, publishes simulation data, and enables discussion could be the solution. Such services are offered by the journal *Geoscientific Model Development* (*9*), but not for the astronomical community.

TRANSPARENCY AND CONSISTENCY

Although citing and sharing are important, the real breakthrough will come with the introduction of a common interface and data-exchange protocol. Very few astronomical simulation codes read each other's data formats or parse each other's input parameters.

This lack of consistency hinders information exchange from one implementation to another and impedes its use by nonfamily researchers. The introduction of a standard for data exchange will eventually lead to more stable software, fewer bugs, and improved validation, verification, and quantification of the results. If only the community could agree on such a standard. Courageous attempts to achieve these objectives include the Virtual Observatory (*10*) for the observational community, Astropy (*11*) for data analysis, and the Astrophysical Multipurpose Software Environment (AMUSE) (*12*) for multiscale simulations. However, the first two of these are not modular, and the third is based on existing community code.

MODULARITY AND TRANSPOSABILITY

One potential analog for a modular astrophysical software system is provided by LEGO bricks, which have a general interface and abstract fine-grained structure that enable endless permutations. For scientific software, however, it is insufficient to design a common interface and establish a modular structure, because the current implementations are too complex to be used in an aggregate environment. Rather than expanding existing code, one can envision the reimplementation of libraries of fundamental solvers that are designed to cooperate. The resulting shorter code has the advantage of being easier to understand, and it contains fewer bugs (*13*). It may seem overkill to rewrite code from scratch, but industry is in a continuous cycle of reengineering.

For general programming, however, the fine-graining in the analogy with LEGO is bound to introduce a variety of optimization problems (*14*). To make the LEGO concept for software design work, the subdivision in minimal routines should not be too rigorous. Instead of dividing the software into its fundamental operations, a more coarse-grained structure could allow collections of fundamental elements to form parts of a larger structure. Twenty years after its introduction, LEGO also realized that many of its young users desire less fine-grained design with more recognizable shapes. The resulting DUPLO blocks reduced graininess while mediating and stimulating experimentation. In terms of software, they would be equivalent to self-contained objects with a common interface structure, which enables inter-object operations.

From a numerical point of view, such an environment should include code-exchange and data-exchange strategies as well as method-coupling paradigms. With the DUPLO analogy, building blocks become the essential ingredient for individual physical domain–specific solvers. These can be easily tested and validated using a daily executed standard test through online services such as Travis CI (*15*). Combinations or communication-essential parts that frequently appear could be reimplemented as separate modules to improve performance and scalability (see the figure). The value per module increases as the environment grows, which again stimulates other researchers to contribute. Independently working dedicated snippets should be publishable and citable—for example, in a code-dedicated journal.

OUTLOOK

We have reached the point where it is no longer reasonable to expect a single graduate student to develop a production-quality astronomical software package. This results in the loss of individual contributions and the further expansion of existing dinosource packages. With a modular approach, a general framework can be complemented by reimplementing dedicated codes that solve only part of the astrophysical problem, each of which can be credited separately. A standardized environment would also stimulate collaboration, mediate validation and verification, expand the scientific life span of individual contributions, and allow the development and maintenance effort to be absorbed by the community. Without the community taking responsibility in establishing a more collaborative environment, the lack of evolution in the engineering of astronomical software will lead to its devolution. ■

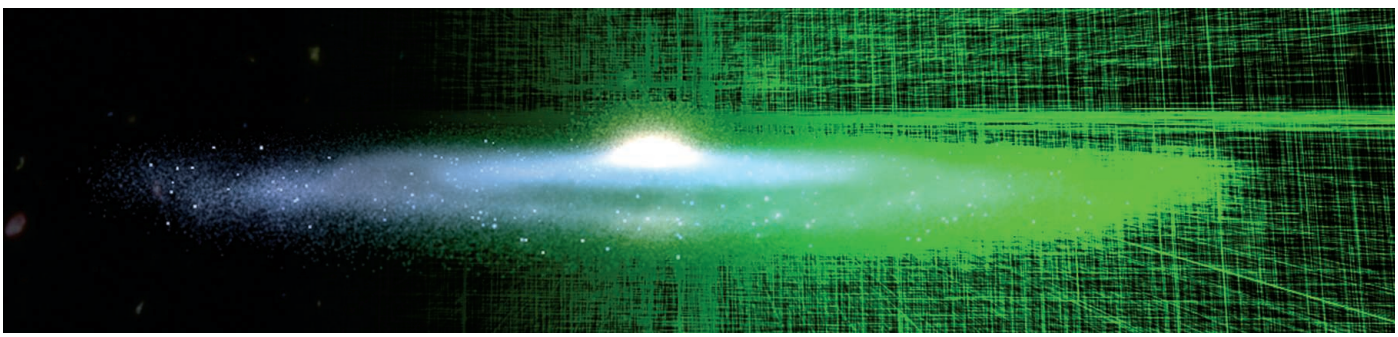


IMAGE: J. BÉDORF *ET AL.* (*16*)

Results from a simulation of the Milky Way Galaxy using the Bonsai code on the ORNL Titan supercomputer (*16*). The left-hand side has been modified to look more like the observations; the right-hand side shows the intrinsic data structure at runtime. Bonsai is a highly optimized but small and dedicated code for a specific application.

REFERENCES AND NOTES

1. S. R. Chidamber, C. F. Kemerer, *IEEE Trans. Softw. Eng.* **20**, 476 (1994).
2. M. M. Lehman, J. F. Ramil, P. D. Wernick, D. E. Perry, W. M. Turski, in *Proceedings of the 4th International Symposium on Software Metrics, METRICS '97* (IEEE Computer Society, 1997); http://dl.acm.org/citation.cfm?id=823454.823901.
3. D. Spinellis, *IEEE/ACM 12th Working Conference on Mining Software Repositories* (2015); 10.1109/MSR.2015.64.
4. F. Bauer, L. Bolliet, H. Helms, in *NATO Software Engineering Conference 1968*, P. Naur, B. Randell, Eds. (Scientific Affairs Division, NATO, 1969).
5. A. Allen *et al.*, *Astrophys. J. Suppl. Ser.* **236**, 10 (2018).
6. C. Collberg, T. A. Proebsting, *Commun. ACM* **59**, 62 (2016).
7. A. Allen *et al.*, *Astron. Soc. Pac. Conf. Ser.* **475**, 387 (2013).
8. Zenodo; https://zenodo.org.
9. *Geoscientific Model Development*; www.geoscientific-model-development.net/.
10. International Virtual Observatory Alliance, *IVOA Architecture Version 1.0: IVOA Note 23 November 2010*; www.ivoa.net/documents/Notes/IVOAArchitecture.
11. Astropy Collaboration, *Astron. Astrophys.* **558**, A33 (2013).
12. S. Portegies Zwart, S. McMillan, *Astrophysical Recipes* (Institute of Physics, 2018).
13. S. McConnell, *Code Complete* (Microsoft Press, ed. 2, 2004).
14. F. P. Brooks Jr., *The Mythical Man-Month* (Addison-Wesley Longman, anniversary ed., 1995).
15. Travis CI; https://travis-ci.org/.
16. J. Bédorf *et al.*, SC14: Proceedings of the International Conference for Hgh Performance Computing, Networking, Storage, and Analysis 10.1109/sc.2014.10 (2014).

ACKNOWLEDGMENTS

I thank J. Bédorf, A. Hoekstra, and A. van Elteren for discussions.

10.1126/science.aau3206

POLICY FORUM

ARMS CONTROL

Denuclearizing North Korea: A verified, phased approach

The process must reflect existing levels of trust at each stage

By **Alexander Glaser** *and* **Zia Mian**

At the June 2018 Singapore Summit, North Korea agreed to the goal of "complete denuclearization" in exchange for "security guarantees" by the United States, including an end to enmity (*1*). Like earlier efforts in the 1990s and 2000s, the current round of diplomacy may well fail because of the challenges of balancing North Korean insistence on the primacy of building trust and cooperation with U.S. demands for progress on denuclearization. Any successful attempt to balance these priorities will have to resolve the thorny question of verification. Here, we propose a phased approach for verified denuclearization that relies on technical measures and tools to allow for the scope, pace, and intrusiveness of denuclearization to reflect progress in political confidence building. More broadly, successfully bridging the goals of denuclearization and political security for North Korea could inform judgments by the international community about how to approach verified disarmament for other states that currently have nuclear weapons.

Although the process of "denuclearization" has not been spelled out explicitly in the current United States–North Korea talks, the two sides seem to have settled on the phrase "complete denuclearization." For the purposes of this analysis, this is taken to include the key nuclear weapon–related obligations agreed on in the 1992 Joint Declaration of South and North Korea on the Denuclearization of the Korean Peninsula, namely to "not test, manufacture, produce, receive, possess, store, deploy or use nuclear weapons" and that these commitments would be verified (*2*).

In March 2018, North Korea announced a moratorium on nuclear weapons and ballistic missile testing. Maintaining this moratorium is seen as the foundation for moving forward with talks and implementing whatever eventually is agreed as the denuclearization process. A more formal commitment to not carry out further nuclear weapon tests would be for North Korea to join the Comprehensive Nuclear Test Ban Treaty (CTBT). Even though the CTBT is not in force, under customary international law, signature of an international treaty confers the obligation to not take actions that would undermine the purpose of the treaty.

Moving forward, eliminating North Korea's nuclear weapons program and related facilities will need a freeze on current weapon-related activities; an agreed baseline of current stockpiles of nuclear weapons, fissile materials, ballistic missiles, and key components; and verified reductions of these stockpiles and downsizing of North Korea's weapons complex. There are already a few proposals for drawn-out, perhaps decade-long, step-by-step approaches that lead to eventual denuclearization, in contrast to demands from Trump administration officials that North Korea "dismantle all of their W.M.D. and ballistic missile programs in a year," but in neither case is attention paid to how verification might assist or hinder such efforts (*3*). To be sustainable, every step in such a process will need to reflect the actual existing level of trust between the United States and North Korea and seek to increase this trust so as to permit future steps. Given the preponderance of U.S. military force, North Korea may set the pace of denuclearization and intrusiveness of the verification measures in case the present process fails, as happened with prior attempts via the 1994 Agreed Framework and the 2003–2009 Six-Party Talks.

We assume that a new framework agreement would contain provisions similar to those in some other arms-control agreements, under which the parties agree not to interfere with specified remote-monitoring techniques or use concealment measures intended to obstruct verification.

FREEZE ON FISSILE MATERIAL

Since North Korea's withdrawal from the Nuclear Non-Proliferation Treaty (NPT) in 2003, there have been essentially no international inspection efforts in North Korea. At the same time, North Korea has expanded the scale and complexity of its nuclear weapons program. On the basis of information available via open sources, it is not clear how many nuclear weapons North Korea possesses today, of what kind (including possibly thermonuclear weapons), and whether they use plutonium or highly enriched uranium (HEU) or both as fissile material. Nor is there reliable information on its ballistic missile capabilities. To establish a basis for moving forward, North Korea could add to its freeze on nuclear weapon and ballistic missile tests a freeze on fissile material production. This can be verified primarily through agreed-on nonintrusive provisions.

Originally, North Korea launched its weapons program with plutonium recovered from the spent fuel of the graphite-moderated (5 MW-electric) plutonium production reactor at Yongbyon. The demolition of its cooling tower in 2008 temporarily made reactor operation impossible and constrained plutonium supply in the following years, but plutonium production at Yongbyon appears to have resumed more recently. In the meantime, North Korea may have shifted the emphasis of its program to uranium enrichment and uranium-based weapons. Today, North Korea most likely produces both plutonium and HEU and may have available material for dozens of nuclear weapons. The question now is how such a freeze could be monitored for both plutonium production and uranium enrichment. North Korea (and South Korea) could permanently refrain from plutonium separation and uranium enrichment, as agreed in their 1992 Joint Declaration.

In the case of plutonium, satellite imagery can be sufficient to confirm the operational status of reactors in North Korea. Imagery can be used to observe heat signatures, vapor plumes, cooling water discharges, and other activities near the reactor (*4*). All these indicators would provide good evidence for a suspension of plutonium production at Yongbyon. Regional krypton-85 monitoring, ideally with a small number of detectors placed around the Yongbyon site, could confirm that remaining spent fuel is not reprocessed (*5*). There are also simple measures to permanently disable the Yongbyon reactor—for example, by blowing boron dust through the core's cooling channels—but North Korea may not agree to such actions until the later stages of the denuclearization process.

The situation with regard to uranium enrichment is more difficult. It may be possible to confirm remotely the shutdown status of the Yongbyon enrichment plant and a possible second plant suspected to

Program on Science and Global Security, Princeton University, Princeton, NJ, USA. Email: alx@princeton.edu

be at Kangson—for example, by monitoring vehicle traffic, including shipments of uranium hexafluoride (UF_6) cylinders entering and leaving the sites, or by monitoring signatures related to electricity supply.

Rather than shut them down, North Korea may prefer to use its enrichment plants for production of low-enriched uranium for its experimental light-water reactor (30 MW-electric). If this or other civilian reactors are allowed to operate, then International Atomic Energy Agency (IAEA) safeguards could be applied to these plants as well as to the feed and product materials associated with them, as happens with civilian uranium enrichment plants in all non-nuclear weapon states and also in some nuclear weapon states. In this case, verification could include unattended measurement systems confirming the nonproduction of HEU, but it would also include onsite inspections. Even if North Korea ended all nuclear activities, IAEA safeguards would still be required to detect possible efforts at reconstitution of its nuclear weapons program.

One major concern is the existence of undeclared nuclear facilities, especially uranium enrichment plants beyond that at Yongbyon and suspected at Kangson. This is a proliferation concern in all states and not limited to North Korea, however (*6*). It is worth noting that in recent decades, undeclared uranium enrichment programs and facilities have been uncovered by using satellite imagery or other reconnaissance and intelligence techniques in Pakistan, Iran, Libya, and North Korea. Big-data analytics and machine-learning techniques will only increase these capabilities over the next decades. For NPT non-nuclear-weapon states, the IAEA safeguards system, strengthened by the measures of the Additional Protocol, seeks to detect such undeclared activities through comprehensive reporting on nuclear activities, including uranium mining and expanded inspection rights. Nonetheless, gaps remain in the ability to detect undeclared enrichment. More broadly, further mechanisms are needed to gain greater confidence in the fact that undeclared nuclear facilities, materials, and items do not exist in a state giving up nuclear weapons. To lay the basis for this process, baseline declarations are needed.

ESTABLISHING A BASELINE

With a freeze as a starting point, declarations of current fissile material and nuclear warhead inventories would be important for measuring progress toward denuclearization. These initial declarations could be relatively simple. Ideally, as a transparency measure, they could be made public. In the case of nuclear warheads, a declaration could include the total number of warheads in North Korea's stockpile, perhaps listed by type, and the number of additional warhead components in storage; in the case of fissile material, a declaration could include acquisitions, losses, and removals, including the aggregate amount of material consumed in tests, and the current inventory of plutonium and highly enriched uranium, ideally also specifying the respective plutonium-239 and uranium-235 contents. More detailed declarations could follow at a later stage of the process.

There is a precedent for fissile material declarations. In May 2008, North Korea declared its plutonium inventory, often reported as 37 kg and backed up by 18,000 pages of operating records. At the time, the United States estimated that North Korea had produced a total of 40 to 50 kg of plutonium, raising concerns that the declaration may be incomplete. U.S. negotiators requested access to the Yongbyon reactor to confirm total plutonium production through use of nuclear archaeological techniques, in which the isotopic ratios of trace impurities in graphite samples are analyzed. At that time, North Korea refused.

Nuclear archaeology techniques for graphite-moderated reactors are now well established and would be sufficient to narrow down the uncertainty in plutonium production to a few kilograms, possibly to less than one weapon-equivalent. North Korea may or may not agree to these procedures early on in the denuclearization process, but every effort must be made to preserve the reactor core and relevant operating records so that such an analysis can be conducted when circumstances permit.

Reconstructing uranium enrichment activities is more challenging. Perhaps the best option would be to reconstruct North Korea's history of uranium supply and use. Such an effort would assess uranium production at North Korean mines, uranium purification, UF_6 production, and enrichment. This would involve auditing the records for internal consistency. Reports of North Korean uranium ore grade suggest that it takes 300 to 400 tons of ore to extract 1 ton of uranium (*7*). This means that up to 2000 tons of ore are required to make 25 kg of weapon-grade HEU or 5 kg of weapon-grade plutonium, the typical amounts used in a nuclear weapon. The review of records from the different plants could be complemented with forensic analysis of tailings at the mines and depleted uranium in cylinders at known enrichment plants. It also may be possible to examine North Korea's centrifuge-plant equipment and reconstruct the amount of uranium processed in these plants and respective HEU output (*8*).

PHOTO: KCNA KCNA/REUTERS

VERIFIED REDUCTIONS

Although establishing baseline declarations of inventories lays an important basis, they do not automatically offer a pathway toward reductions of nuclear weapons and fissile materials. Here, we briefly outline three possible strategies to set such a process in motion. Any one of them or some agreed combination could offer a viable path forward. Because the denuclearization process can be expected to be fragile, especially in the beginning, when the mutual level of distrust will be particularly high, it is important to distinguish reasonable short-term confidence-building steps from more ambitious longer-term goals that involve actual denuclearization steps and the related verification measures. Regardless of the path taken, it will take years to conclude that undeclared stockpiles of materials and warheads do not exist, even if North Korea fully cooperates.

As a first option, while providing baseline information on the size of its stockpiles, North Korea may initially not want to reveal the storage locations of its nuclear warheads, warhead components, and long-range missiles. Such items could, however, still be monitored while in storage as a major confidence-building measure. Warheads, for example, could be brought from their storage location to some agreed sites and placed

Kim Jong-un (center) stands next to a nuclear warhead in a photo released the day before North Korea's sixth nuclear test on 3 September 2017.

in tagged containers with agreed-on special electronic or optical seals (*9*). For each warhead, radiation measurements could confirm that items declared as nuclear warheads or components have signatures that are consistent with items containing kilogram quantities of plutonium and uranium. Warheads are then returned to secret storage locations (*10*). Periodically, North Korean personnel could prove the integrity of seals and containers by presenting randomly selected items again for a physical inspection.

A second option could be for North Korea simply to deliver nuclear weapons to be dismantled in the presence of inspectors at an agreed facility or submit a specified amount of plutonium and highly enriched uranium recovered from dismantled warheads for international safeguarding. For safety reasons, as former Los Alamos National Laboratory Director Siegfried Hecker and colleagues recently observed, "shipping the North's nuclear weapons out of the country is naïve and dangerous. The weapons must be disassembled by the people who assembled them" (*11*). Eventually, the amount of fissile material submitted would have to match the amount of material declared to have been produced and used for weapons (*12*).

A third option would be for North Korea to gradually reduce the size of its weapons complex without revealing where exactly nuclear weapons and long-range ballistic missiles remain. An estimate in 2014 suggested about 90 nuclear weapon and missile sites of potential interest (*13*). North Korea's declarations in this approach could include locations of production, storage and deployment sites, description of the facilities at each site, and relevant operating records. The data need not be revealed all at once if North Korea was concerned about providing a potential target list to the United States. North Korea could provide such a declaration in hashed form and reveal information only when required for the inspection of a particular site (*14*). It would only offer for inspection sites that have been "cleaned out" and could keep the most important sites for last. This plan builds on the concept of "deferred verification" developed at the United Nations Institute for Disarmament Research (*15*). This approach envisions a state's nuclear enterprise as including a closed segment, where all military activities take place, and an expanding open segment that is fully accessible to inspectors. For the United States, it serves to create a predictable path in which North Korea commits in advance to a step-by-step process of declaring all nuclear sites, activities, and materials for inspection or to confirm their elimination. This concept aims to give North Korea flexibility because it might seek to maintain some nuclear capabilities until it is confident that the peace process has matured.

The elimination of North Korea's weapons program and reestablishing its formal status as a non-nuclear-weapon state will need a legal and institutional structure. This could be a North Korean return to the NPT. An additional new possibility is accession to the 2017 Treaty on the Prohibition of Nuclear Weapons, which requires parties to commit to a "time-bound plan for the verified and irreversible elimination of that State Party's nuclear-weapon programme, including the elimination or irreversible conversion of all nuclear-weapons-related facilities."

The current efforts to resolve the North Korean nuclear crisis are an uncertain but potentially historic opportunity to eliminate a nuclear weapons program. They can also offer insights into the potential critical technical and political obstacles involved in tackling the arsenals of other nuclear weapon states. ■

REFERENCES AND NOTES

1. Joint Statement of President Donald J. Trump of the United States of America and Chairman Kim Jong Un of the Democratic People's Republic of Korea at the Singapore Summit, 12 June 2018.
2. Joint Declaration of the Denuclearization of the Korean Peninsula, 20 January 1992.
3. D. E. Sanger, W. J. Broad, "As Bolton says North Korea could disarm in a year, reality lags promises," *New York Times*, 1 July 2018.
4. D. Albright *et al.*, *Ann. Rev. Earth Planet. Sci.* **46**, 99 (2018).
5. M. Schöppner, *Science Glob. Sec.* **26**, 1 (2018).
6. R. S. Kemp, *Ann. Rev. Earth Planet. Sci.* **44**, 17 (2016).
7. J. Lewis, "Recent imagery suggests increased uranium production in North Korea, probably for expanding nuclear weapons stockpile and reactor fuel," 38 North, 12 August 2015; www.38north.org/2015/08/jlewis081215.
8. T. W. Wood *et al.*, *Sci. Glob. Sec.* **22**, 4 (2014).
9. Relevant concepts could use, for example, hardware tokens that display periodically updated pseudo-random numbers. The fact that the host can still report the current number would prove the integrity of the seal and the container, even if inspectors are not present for the readout of the token. A similar concept is described at share-ng.sandia.gov/news/resources/news_releases/tamper_detection .
10. Decoy containers and vehicles could be used to obfuscate true routes and destinations. Alternatively, inspections could take place at selected portals to the extensive tunnel networks that are believed to exist in North Korea and that may hold its nuclear weapons.
11. S. S. Hecker *et al.*, *A Technically-informed Roadmap for North Korea's Denuclearization* (Center for International Security and Cooperation, Stanford University, 28 May 2018).
12. The amounts of fissile material consumed in nuclear tests would be particularly difficult to independently verify. As part of a denuclearization agreement, North Korea could declare the amounts of fissile material consumed in each test. Experts from weapon states could review these and other data North Korea may provide to verify the correctness of their declaration; perhaps, this could also include the design of the test devices. Last, as an ultimate measure if discrepancies cannot be resolved, one could drill into the test shafts to obtain samples that could be used for a postdetonation forensic analysis.
13. M. J. Mazarr *et al.*, *The Korean Peninsula: Three Dangerous Scenarios* (RAND, 2018); www.rand.org/content/dam/rand/pubs/perspectives/PE200/PE262/RAND_PE262.pdf.
14. S. Philippe *et al.* (2017), "Cryptographic Escrow of Nuclear Warhead Inventories for Early Commitment and Non-intrusive Verification," *Proc. 58th INMM Annual Meeting*, 16 to 20 July 2017, Indian Wells, California, USA.
15. P. Podvig, J. Rodgers, *Deferred Verification: Verifiable Declarations of Fissile Material Stocks*, UNIDIR 2017; unidir.org/files/publications/pdfs/deferred-verification-verifiable-declarations-of-fissile-material-stocks-en-694.pdf.

10.1126/science.aau4817

AAAS Leshner Leadership Institute fellows are calling on AAAS to instate a harassment policy that rescinds honors in the case of misconduct.

Edited by **Jennifer Sills**

Call for new AAAS harassment policy

As fellows of the AAAS Leshner Leadership Institute for Public Engagement with Science ("Public engagement helps scientists tackle global challenges," A. Q. Hoy, Association Affairs, 27 July, p. 372), we are writing to express concern that AAAS (the publisher of *Science*) continues to honor scientists who have engaged in harassment. Harassment, including sexual and gender harassment, is scientific misconduct, and its effects influence the daily lives of scientists, especially those who are from underrepresented populations. Honoring harassers sends a message to the entire scientific community that a harasser's individual scientific achievements are considered more valuable than their victims as well as more valuable than the severe, widespread effects of a culture of harassment on the careers, livelihoods, and scientific potential of a much broader population.

Currently, no clear mechanism exists for preventing individuals who have damaged science by engaging in harassment from receiving and retaining awards, titles, and honors from AAAS. We urge AAAS to adopt a strong, enforceable policy to address harassment and discrimination, along with other types of scientific misconduct, among honorees (including elected fellows and award recipients). A meaningful policy would be applied prospectively and retroactively, require disclosure of institutional findings of professional misconduct, establish transparent and accountable procedures to report incidents, and make recommendations for appropriate responses, including revoking honors. Such an awards and honors policy would send a message that behavior that harms, degrades, and discriminates is incompatible with attaining the highest levels of scientific recognition. We would welcome its application to our own and future Leshner cohorts.

We were selected as Leshner fellows for our commitment to public engagement and institutional change. An inclusive scientific community is necessary for fulfilling these goals. By taking a leadership role on this important issue, we believe that AAAS can serve as a model for other institutions and professional societies.

Noelle E. Selin,[1]* Melissa A. Kenney,[2] Anne J. Jefferson,[3] Jeffrey S. Dukes,[4] Tessa M. Hill,[5] Laura Schmitt Olabisi,[6] Meghan A. Duffy[7]

[1]Institute for Data, Systems, and Society and Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]CMNS-Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD 20740, USA. [3]Department of Geology, Kent State University, Kent, OH 44242, USA. [4]Department of Forestry and Natural Resources and Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA. [5]Department of Earth and Planetary Sciences, University of California, Davis, Davis, CA 95616, USA. [6]Department of Community Sustainability, Michigan State University, East Lansing, MI 48823, USA. [7]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA.
*Corresponding author. Email: selin@mit.edu

NOTE

1. An expanded version of our letter and a list of signatories is available at https://sites.google.com/view/timesupaaas.

10.1126/science.aav1680

Improving societies' harassment policies

As Howard Hughes Medical Institute (HHMI) professors, teachers, and mentors, we are acutely aware of the harm that can be done by sexual harassment and other discriminatory behaviors, which negatively affect the careers of young scientists and hamper our efforts to diversify the scientific workforce and professoriate. We applaud the recent report on sexual harassment of women, climate, and culture from the National Academies of Sciences, Engineering, and Medicine and the recommendations therein (*1*). Scientific societies can play an important role in changing discriminatory culture.

A number of societies have recently implemented or improved codes of conduct. The American Geophysical Union includes harassment as a form of scientific misconduct under its new ethics policy (*2*, *3*). The Society of HHMI Professors has recently changed its policies so that

membership, either initial or continuing, requires that the person be in good standing at their university or other employer in terms of relevant codes of conduct (*4*). AAAS (the publisher of *Science*) is in the process of writing a new policy (*5*), and there is a petition urging the U.S. National Academy of Sciences to enact such policies as well (*5*). We encourage all professional societies to do likewise. These steps, if taken by all, will help the scientific community to maintain respect for all members, foster an environment of inclusion, and support a diverse workforce.

Ariel Anbar,[1] Sarah Elgin,[2] Joseph Jez,[2] Diane O'Dowd,[3] Beth Shapiro,[4] Muhammad Zaman[5]*

[1]School of Earth and Space Exploration and School of Molecular Sciences, Arizona State University, Tempe, AZ 85287, USA. [2]Department of Biology, Washington University in St. Louis, St. Louis, MO 63130, USA. [3]Department of Developmental and Cell Biology, University of California, Irvine, Irvine, CA 92697, USA. [4]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA 95064, USA. [5]Departments of Biomedical Engineering and International Health, Boston University, Boston, MA 02215, USA.
*Corresponding author. Email: zaman@bu.edu

REFERENCES AND NOTES

1. National Academies of Sciences, Engineering, and Medicine, "Sexual harassment of women: Climate, culture, and consequences in academic sciences, engineering, and medicine" (The National Academies Press, Washington, DC, 2018).
2. M. J. McPhaden, L. Gundersen, B. M. Williams, *Eos* **98**, 10.1029/2017EO082469 (2017),
3. E. Davidson, "AGU's revised ethics policy: Where we are 6 months later," *From the Prow* (2018); https://fromtheprow.agu.org/agus-revised-ethics-policy-where-we-are-6-months-later/.
4. The Society of HHMI Professors, Mission Statement and Rules of Order (2018); www.hhmi.org/sites/default/files/society_of_hhmi_professors_mission_and_rules_of_order.approved_2018.i2.pdf.
5. M. Wadman, *Science* **360**, 949 (2018).
6. The full list of signatories is available as supplementary material.

SUPPLEMENTARY MATERIALS

Full List of Signatories
www.sciencemag.org/content/361/6406/984.2/suppl/DC1

10.1126/science.aav1362

ERRATA

Erratum for the Report "Synapse-specific representation of the identity of overlapping memory engrams" by K. Abdou *et al.*, *Science* **361**, eaau8829 (2018). Published online 3 August 2018; 10.1126/science.aau8829

Erratum for the Report "Aging and neurodegeneration are associated with increased mutations in single human neurons" by M. A. Lodato *et al.*, *Science* **361**, eaau6185 (2018). Published online 6 July 2018; 10.1126/science.aau6185

OUTSIDE THE TOWER

Science engagement in South Africa

Learners from grades 9 to 12 surrounded our Science Week table in the township Khaye-litsha, an impoverished community near Cape Town in South Africa. We were conducting outreach for our project Cape Citizen Science (http://citsci.co.za/), an initiative to engage nonscientists in plant disease research in a global biodiversity hotspot. "Can plants get sick, too?" we asked, as the students examined unhealthy plants under dissecting microscopes and held Petri plates containing fungal-like organisms up toward the lights. Late in the day, a grade 10 boy named Dylan approached the table and began asking about our work. His hunger to learn and the depth of his questions inspired us to invite him to our lab. He immediately responded, "Can I bring my friends?"

Many researchers in our community met Dylan, Ayebonga, and Ivan during the next year, as they often joined us in our lab to satisfy their curiosity and contribute their time to the scientific process. Their dedication to learning was exemplified by the challenges they overcame to travel to our university; they often spent hours navigating the public transit system of the Western Cape Province and occasionally faced financial barriers (although we made sure to reimburse them for their efforts).

Call for Submissions
Outside the Tower is an occasional feature highlighting scientists' advocacy experiences. Submit your advocacy story at http://cts.sciencemag.org.

On his first visit to the lab, Dylan excitedly informed us that it was his first time looking through a microscope. After a few visits, he told us that he wants to study microbiology at university. We are grateful to know that we helped empower him to make such a critical decision. However, there are thousands of underprivileged learners in South Africa without these opportunities, whose hunger for knowledge remains unfulfilled and overlooked. We encourage researchers to participate in public engagement programs, especially in countries with emerging economies. Such programs can increase the value and capacity of research beyond science. We also encourage leaders of traditional science projects to open their programs to citizens, especially those without reliable access to quality science education. Together, we can help science enthusiasts (and potential future scientists) develop critical decision-making and problem-solving skills.

J. M. Hulbert[1,2]* and F. Roets[2]

[1]Forestry and Agricultural Biotechnology Institute, University of Pretoria, Hatfield, 0028, South Africa. [2]Department of Conservation Ecology and Entomology, Stellenbosch University, Matieland, 7602, South Africa.
*Corresponding author. Email: joey.hulbert@fabi.up.ac.za

10.1126/science.aav1499

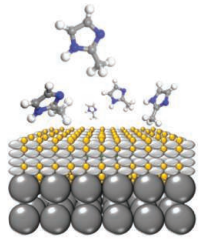
High school students visit a lab after meeting scientists at a public engagement program.

PHOTO: AIMS-SOUTH AFRICA

RESEARCH

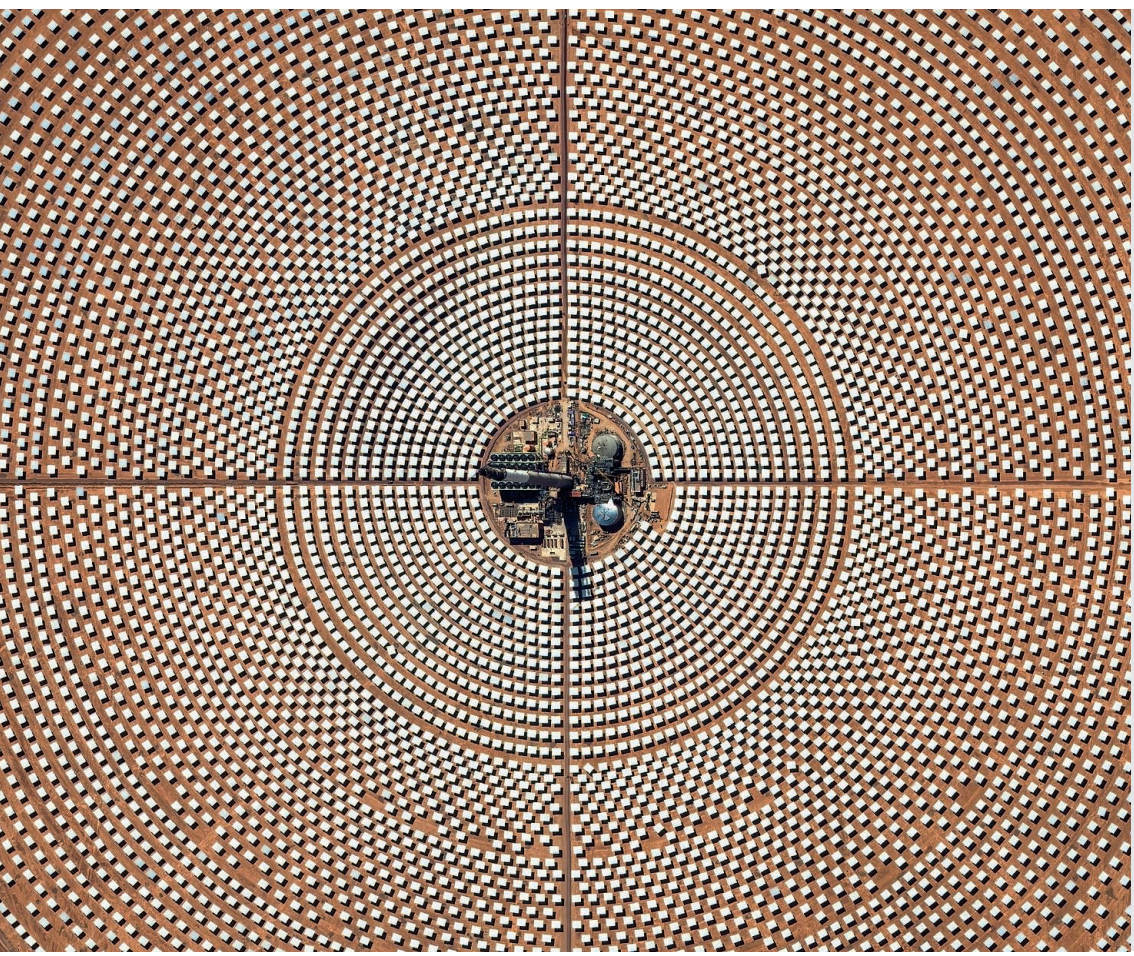
Zeolitic imidazolate framework membranes

Ma et al., *p. 1008*



IN *SCIENCE* JOURNALS

Edited by **Stella Hurtley**



CREDITS: (TOP TO BOTTOM) MA *ET AL.*: DIGITALGLOBE/SCAPEWARE3D/CONTRIBUTOR/GETTY IMAGES

ENERGY

More energy, more rain

Energy generation by wind and solar farms could reduce carbon emissions and thus mitigate anthropogenic climate change. But is this its only benefit? Li *et al*. conducted experiments using a climate model to show that the installation of large-scale wind and solar power generation facilities in the Sahara could cause more local rainfall, particularly in the neighboring Sahel region. This effect, caused by a combination of increased surface drag and reduced albedo, could increase coverage by vegetation, creating a positive feedback that would further increase rainfall. —HJS

Science, this issue p. 1019

Large-scale solar power generation facilities in the Sahara could cause more local rainfall.

OPTICAL COMPUTING

All-optical deep learning

Deep learning uses multilayered artificial neural networks to learn digitally from large datasets. It then performs advanced identification and classification tasks. To date, these multilayered neural networks have been implemented on a computer. Lin *et al*. demonstrate all-optical machine learning that uses passive optical components that can be patterned and fabricated with 3D-printing. Their hardware approach comprises stacked layers of diffractive optical elements analogous to an artificial neural network that can be trained to execute complex functions at the speed of light. —ISO

Science, this issue p. 1004

GALAXY FEEDBACK

Molecular gas ejected from a distant galaxy

Galaxies grow by forming stars from cold molecular gas. The rate at which they do so is limited by various feedback processes (such as supernovae or stellar winds) that heat and/or eject gas from the host galaxy. Spilker *et al*. used submillimeter observations to discover an outflow of molecular gas from a galaxy in the early Universe, a period of vigorous star formation. Modeling the outflow revealed that the mass of gas being ejected is similar to that being turned into stars. The results will help determine how quickly galaxies formed after the Big Bang. —KTS

Science, this issue p. 1016

MIGRATION

Learning where and when

Large ungulate migrations occur across continents and inspire curiosity about how these animals know when to leave and where to go. Jesmer *et al*. took advantage of regional extinctions and reintroductions of several North American ungulate species to determine the role of learning in migrations (see the Perspective by Festa-Bianchet). Reintroduced populations of bighorn sheep and moose did

Published by AAAS

not migrate as historical herds had. However, after several decades, newly established herds were better able to track the emergence of vegetation in the environment and were increasingly migratory. Thus, newly introduced animals learned about their environment and shared the information through social exchange. —SNV

Science, this issue p. 1023; see also p. 972

FERTILIZATION

Bouncer keeps fertilization specific

Fertilization needs to be highly efficient while remaining species-specific. However, despite decades of research, it is still unclear how these two requirements are met. Herberg *et al*. report the discovery of the Ly6/uPAR-type protein Bouncer as a species-specific fertilization factor in zebrafish (see the Perspective by Lehmann). Bouncer localizes to the egg membrane and is required for sperm entry. Remarkably, expression of Bouncer from another fish species (medaka) in zebrafish allowed for cross-species fertilization. —BAP

Science, this issue p. 1029; see also p. 974

CANCER

Metastatic drivers same as primary

Treatment decisions for cancer patients are increasingly guided by analysis of the gene mutations that drive primary tumor growth. Relatively little is known about driver gene mutations in metastases, which cause most cancer-related deaths. Reiter *et al*. explored whether the growth of different metastatic lesions within an individual patient is fueled by the same or distinct gene mutations. In a study of 76 untreated metastases from 20 patients with different types of cancer, all metastases within a patient shared the same functional driver gene mutations. Thus, analysis of a single biopsy could help oncologists select the optimal therapy for patients with widespread metastatic disease. —PAK

Science, this issue p. 1033

ANTHROPOLOGY

Organization of historical cemeteries

The organization of Early Medieval cemeteries of the Alemanni, a Germanic tribe, is thought to be based on households. However, specific kinship relationships between individuals found in these cemeteries have not been tested formally. O'Sullivan *et al*. examined 13 individuals from the Niederstotzingen cemetery in southern Germany using ancient DNA. The strontium and oxygen isotope content of their dental enamel revealed that, whereas five of the individuals were second-degree relatives born in the region, two others were of nonlocal origin. Thus, other social processes, such as personal fealty to powerful families, might also have influenced the composition of these cemeteries. —MSA

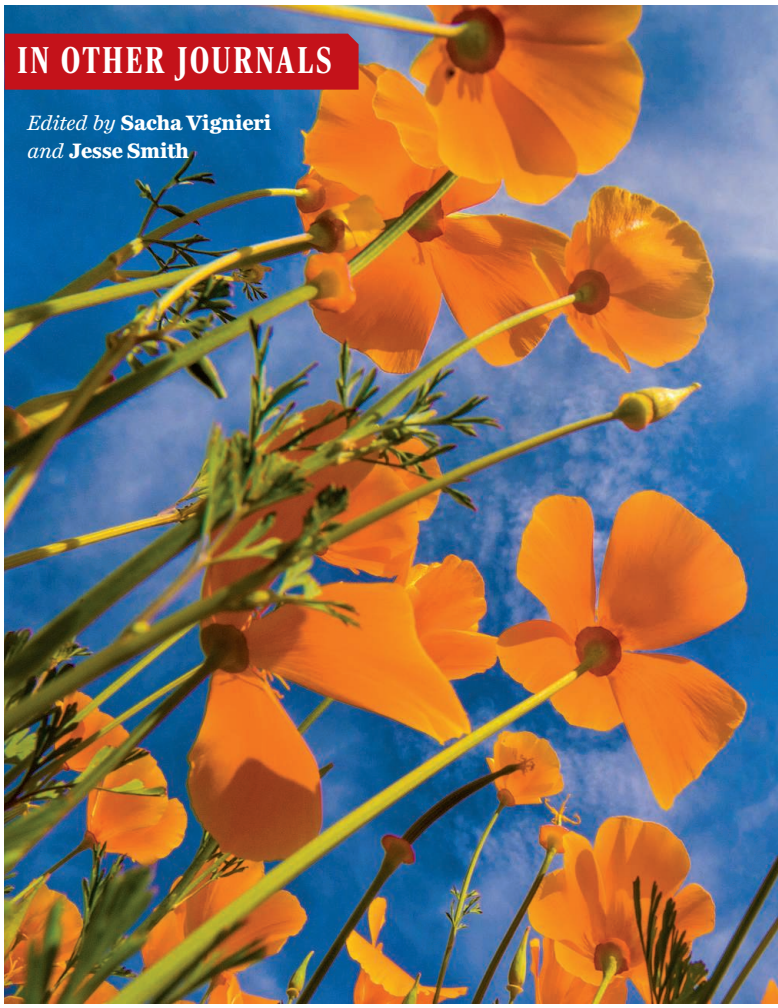
Sci. Adv. 10.1126/sciadv.aao1262 (2018).

INFECTIOUS DISEASES

Killer trifecta for leprosy

Cytotoxic granule proteins secreted by $CD8^+$ T cells assist in the killing of both infected cells and intracellular bacteria. Balin *et al*. studied the ability of subsets of human $CD8^+$ T cells that express different combinations of granule proteins to kill macrophages infected with *Mycobacterium leprae*. The $CD8^+$ T cell subset with the highest efficiency of mycobacterial killing simultaneously expressed three granule proteins: granzyme B, perforin, and granulysin. Transcriptional profiling of $CD8^+$ T cell subsets identified the natural killer (NK)–activating receptor NKG2C as a surface marker for identification and enrichment of these potent antimicrobial cells. —IW

Sci. Immunol. **3**, eaat7668 (2018).



IN OTHER JOURNALS

Edited by **Sacha Vignieri** *and* **Jesse Smith**

SYNTHETIC BIOLOGY

Synthesizing a therapeutic probiotic

Phenylketonuria is a disease caused by an inability to metabolize the amino acid phenylalanine, which can accumulate in the blood and brain, causing neurotoxicity. Patients are treated by restricting phenylalanine intake through a low-protein diet, but this can cause failure to thrive. To improve the therapeutic options, Isabella *et al*. developed a probiotic that meets the current requirements for clinical testing. They engineered a strain of *Escherichia coli* with a strong safety profile in humans to inducibly express a phenylalanine-degrading enzyme. Oral administration of this probiotic in a mouse model of phenylketonuria prevented increased phenylalanine concentrations in the blood when the mice were injected with phenylalanine, suggesting that gastrointestinal degradation can regulate circulating phenylalanine concentrations. Thus, this synthetic probiotic could have potential in clinical trials. —GKA

Nat. Biotechnol. 10.1038/nbt.4222 (2018).

CANCER

Watching kidney cancer metabolism

Clear cell renal cell carcinoma (ccRCC) is the most common and aggressive form of kidney cancer and undergoes extensive metabolic reprogramming. Courtney *et al*. infused a glucose isotope into patients with primary ccRCC who were undergoing surgery and traced metabolic and isotopic flux. Compared with cells of the adjacent kidney, tumor cells exhibited prominent

PHOTO: TIM FITZHARRIS/MINDEN PICTURES/GETTY IMAGES

PLANT SCIENCE

Cell wall microlensing produces brilliant flower color

The California poppy is a drought-tolerant plant with brilliant yellow to orange flowers. Wilts *et al*. looked more closely to reveal the optics behind these intense colors. The surface of the flower's petals is covered with a tidy array of microscopic ridges that act like prisms. The ridges develop with successive deposits of cell wall and are aligned by rows of cells. When the ridge has been built and the cell has reached optimal size, light captured by the prism-like ridge is focused by microlensing onto carotenoid pigment granules at the bottom of the cell, revealing the mature flower's color. Light from certain angles is reflected back off the edges of the prisms, contributing to the silky appearance of this flexible petal. A little bit of pigment goes a long way when light delivery is optimized. —PJH

New Phytol. **219**, 1124 (2018).

Microscopic ridges enhance California poppy color through prismatic effects.

PHOTO: RHONDA GUTENBERG/GETTY IMAGES

glycolysis, whereas the presence of tricarboxylic acid (TCA) cycle metabolites (indicating glucose oxidation) was diminished. ccRCC tumors were more glycolytic compared with brain and lung tumors from different patients. In one patient who was infused with an acetate isotope (acetate is a direct substrate of the TCA cycle), low TCA cycle turnover of metabolites was also observed. This phenomenon describes the Warburg effect of metabolism in ccRCC and highlights metabolic differences between different types of cancer. —MY

Cell Metab. 10.1016/j.cmet.2018.07.020 (2018).

QUANTUM OPTICS

A quantum optical storage box

The development of information technologies based on quantum optics relies crucially on the ability to generate and propagate single photons in free space or optic fibers over vast distances. However, information processing also requires memory or storage functionality, and the ability to hold on to photons tends to pose a challenge. Seri *et al*. used femtosecond laser pulses to micromachine waveguides in an optical crystal of $Pr^{3+}:Y_2SiO_5$. By identifying irradiation conditions that avoid damage to the crystal, they found that the properties of the waveguide, such as the storage time for single photons, are enhanced. Relatively long storage times, combined with compatibility with other optical processing technologies, could provide a powerful platform for developing complex integrated quantum architectures. —ISO

Optica **5**, 934 (2018).

HUMAN GENETICS

Linking disease genotypes to phenotype

The penetrance of a genetic variant is the degree to which a specific genetic change affects an individual's phenotype. However, it is not clear why a specific pathogenic mutation exhibits an unpredictable phenotype among individuals. Castel *et al*. examined the genomes and expression of RNA across individuals and found that deleterious mutations affecting protein-coding genes are more likely to be linked to regulatory elements that lower the expression of the pathogenic gene—hence lowering the overall penetrance of the mutation. However, relative to unaffected people, the overall penetrance was higher in individuals with cancer and autism. These results suggest that the joint effects of regulatory and coding mutations are subject to purifying selection to reduce penetrance. —LMZ

Nat. Genet. 10.1038/s41588-018-0192-y (2018).

METALLURGY

Tracking corroding chloride

Corrosion is a major problem for metals, reducing their performance and shortening their lifetimes. Passive films that form on metal surfaces can help provide corrosion resistance but are also susceptible to localized attack. Zhang *et al*. used transmission electron microscopy to take a detailed look at chloride attack on the passive films of an iron-chromium-nickel alloy. This direct visualization of chloride-ion transport allows for a much better understanding of how these ions modify the surface layer of metals and deeper insight into the corrosion process. —BG

Nat. Commun. 10.1038/s41467-018-04942-x (2018).

SOCIAL NETWORKS

Segregated travel patterns within cities

People within cities are often segregated by race and class. To find out if segregation is limited to where people reside or if it extends to movements outside of their homes, Wang *et al*. analyzed 650 million geocoded tweets from 400,000 residents of America's 50 most populous cities to track their travel patterns. They found that people who live in primarily black or Hispanic neighborhoods, regardless of income, were less likely to travel to white or middle-class neighborhoods, even more so than residents of poor white neighborhoods. This relationship held even though all groups travel approximately the same distance during their day to the same number of neighborhoods. These data have implications for understanding the breadth of racial segregation within cities. —TSR

Proc. Natl. Acad. Sci. U.S.A. **115**, 7735 (2018).

Chloride ions attack passive films on metal surfaces and cause corrosion.

ALSO IN *SCIENCE* JOURNALS

Edited by **Stella Hurtley**

ECOLOGY

Making sense of transient dynamics

Ecological systems can switch between alternative dynamic states. For example, the species composition of the community can change or nutrient dynamics can shift, even if there is little or no change in underlying environmental conditions. Such switches can be abrupt or more gradual, and a growing number of studies examine the transient dynamics between one state and another—particularly in the context of anthropogenic global change. Hastings *et al*. review current knowledge of transient dynamics, showing that hitherto idiosyncratic and individual patterns can be classified into a coherent framework, with important general lessons and directions for future study. —AMS

Science, this issue p. 990

NEURODEGENERATION

Adult neurogenesis and Alzheimer's disease

Alzheimer's disease (AD) pathology destroys neurons and synapses in the brain, leading to dementia. The brain generates new neurons throughout life in the hippocampus, a process called adult hippocampal neurogenesis (AHN). Choi *et al*. found that blocking AHN exacerbated cognitive impairment in an AD mouse model (see the Perspective by Spires-Jones and Ritchie). Inducing neurogenesis alone did not improve cognition in AD mice, whereas inducing neurogenesis while simultaneously ameliorating the neuronal environment via exercise did. The use of genetic or pharmacological treatments that simultaneously induced neurogenesis and increased levels of brain-derived neurotrophic factor (BDNF) mimicked the benefits of exercise on cognition. Thus, inducing both neurogenesis and providing BDNF may be useful as an AD therapeutic. —SMH

Science, this issue p. 991; see also p. 975

STRUCTURAL BIOLOGY

A complex implicated in kidney health

Autosomal dominant polycystic kidney disease (ADPKD) is a common genetic disease that can lead to kidney failure. Mutations in the proteins PKD1 and PKD2 are linked to the disease, but the function of these proteins remains unclear, both in physiology and disease. PKD1 has been implicated in the sensing of chemical and mechanical force stimuli, and PKD2 is proposed to be a calcium ion channel. Su *et al*. show that the transmembrane regions form a PKD1-PKD2 complex assembled in a 1:3 ratio. Their high-resolution cryo–electron microscopy structure confirms that the complex adopts transient receptor potential channel architecture, with some distinctive features. Mapping disease-causing mutations onto the structure suggests that pathogenesis may come from incorrect folding or trafficking of the complex rather than from disruption of channel activity. —VV

Science, this issue p. 992

TOPOLOGICAL OPTICS

Generating a lattice of optical skyrmions

The topological properties of systems can be robust to defects and imperfections. Skyrmions are one such topological entity, which resemble "hedgehog-like" structures. Tsesses *et al*. controlled the interference of plasmon polaritons on a patterned metallic surface to generate a lattice of optical skyrmions. Such control could potentially optically mimic complex condensed matter systems or help in the development of robust optical communications and circuits. —ISO

Science, this issue p. 993

MEMBRANES

The makings of permeable membranes

A challenge in making membranes is finding ways to control the pore structure during the fabrication process or through postfabrication treatment. The deposition of a zeolitic imidazolate framework material called ZIF-8 onto an alumina support gives a dense, impermeable material. However, when Ma *et al*. exposed this material to vapors of 2-methylimidazole, it transformed into a porous material able to separate propylene from propane. —MSL

Science, this issue p. 1008

SURFACE CHEMISTRY

Selectively exciting desorption

Reactions of molecules adsorbed on surfaces can be induced by injecting electrons from the tip of a scanning tunneling microscope. Rusimova *et al*. show that for the tip-induced desorption of toluene molecules from a silicon surface, two activation channels exist: One is invariant, but the other depends on the height of the tip above the surface. When the tip is very close to the molecule, it can quench the excitation. The decreased lifetime, in turn, decreases the desorption probability. —PDS

Science, this issue p. 1012

PLANT SCIENCE

Flexible growth and immune responses in rice

Plants that are fighting microbial pathogens often divert resources that could be used for growth into the immune response. For crops, this translates into lower yield when plant immunity is activated. Wang *et al*. show that, in rice, reversible phosphorylation of a key transcription factor allows the plant to defend against fungal attack when needed but then, within days, reallocate resources back to growth (see the Perspective by Greene and Dong). Thus, both pathogen defense and crop yield can be sustained. —PJH

Science, this issue p. 1026; see also p. 976

PLASMA ASTROPHYSICS

Two-step energy transfer in space plasma

Plasmas are ionized gases that contain negative electrons, positive ions, and electromagnetic fields. These constituents can oscillate in position over time, carrying energy as plasma waves. In principle, such waves could transfer energy between two different ion populations. Kitamura *et al*. analyzed data from the Magnetospheric Multiscale mission, a group of four spacecraft that are flying in tight formation through Earth's magnetosphere. They discovered an event in which energy was transferred from hydrogen ions to plasma waves and then from the waves to helium ions. This energy transfer process is likely to occur in many other plasma environments. —KTS

Science, this issue p. 1000

ASTROPHYSICS

A modular approach to astrophysics

Astrophysical modeling is essential for furthering the understanding of our Universe, from star formation to the evolution of galaxies. In a Perspective, Portegies Zwart argues that the software used for this modeling is not fit for purpose. Although astrophysical simulation codes are relatively simple in structure,

a lack of standards means that few codes are compatible with each other and that each code exists as a monolith that is only understandable to the developers. A modular framework of open-source, citable code units that are compatible with each other could address these problems. Such a framework would increase the recognition of individual contributions, particularly from young scientists, and would further astrophysical discovery. —JFU

Science, this issue p. 979

COMBUSTION CHEMISTRY

A radical route to soot

The chemical origin of soot is a persistent puzzle. It is clear that small hydrocarbon fragments formed in flames must aggregate into larger particles, but the initial driving force for aggregation remains a mystery. Johansson *et al*. combined theory and mass spectrometry to suggest a solution based on resonance-stabilized radicals (see the Perspective by Thomson and Mitra). Aromatics such as cyclopentadiene have a characteristically weak C–H bond because their cleavage produces radicals with extended spans of π-electron conjugation. Clusters thus build up through successive coupling reactions that extend conjugation in stabilized radicals of larger and larger size. —JSY

Science, this issue p. 997; see also p. 978

STRESS RESPONSES

The right proteins at the right time

Cells must tailor protein synthesis so that proteins that enable growth and proliferation and those that enable adaptation under stressful conditions are produced at the appropriate times. Torrent *et al*. found that yeast genes encoding proteins involved in growth and proliferation used common codons, whereas those encoding stress-response proteins tended to use rare codons (see the Focus by Pechmann). Stress skewed the transfer RNA (tRNA) pool toward those tRNAs that recognize rare codons. Furthermore, stress proteins were translated faster than those involved in growth and proliferation. —WW

Sci. Signal. **11**, eaat6409, eaau1098 (2018).

ASTHMA

Remodeling airway innervation in asthma

Asthma is a widespread chronic airway disease characterized by airway obstruction, inflammation, and hyperresponsiveness. In eosinophilic asthma—the most common form of asthma—eosinophils in the airway alter nerve function and exacerbate the disease. Drake *et al*. studied samples obtained by endobronchial biopsy from patients with severe eosinophilic asthma. They found that airway innervation was increased and positively correlated with symptom severity. In mice, eosinophilia increased airway innervation and triggered bronchoconstriction and airway hyperresponsiveness. Thus, structural remodeling of airway innervation contributes to symptom severity in eosinophilic asthma. —MM

Sci. Transl. Med. **10**, eaar8477 (2018).

REVIEW SUMMARY

ECOLOGY

Transient phenomena in ecology

Alan Hastings*, Karen C. Abbott, Kim Cuddington, Tessa Francis, Gabriel Gellner, Ying-Cheng Lai, Andrew Morozov, Sergei Petrovskii, Katherine Scranton, Mary Lou Zeeman

BACKGROUND: Much of ecological theory and the understanding of ecological systems has been based on the idea that the observed states and dynamics of ecological systems can be represented by stable asymptotic behavior of models describing these systems. Beginning with early work by Lotka and Volterra through the seminal work of May in the 1970s, this view has dominated much of ecological thinking, although concepts such as the idea of tipping points in ecological systems have played an increasingly important role. In contrast to the implied long time scales of asymptotic behavior in mathematical models, both observations of ecological systems and questions related to the management of ecological systems are typically focused on relatively short time scales.

A number of models and observations demonstrate possible transient behavior that may persist over very long time periods, followed by rapid changes in dynamics. In these examples, focusing solely on the long-term behavior of systems would be misleading. A long transient is a persistent dynamical regime—including near-constant dynamics, cyclic dynamics, or even apparently chaotic dynamics—that persists for more than a few and as many as tens of generations, but which is not the stable long-term dynamic that would eventually occur. These examples have demonstrated the potential importance of transients but have often appeared to be a set of idiosyncratic cases. What is needed is an organized approach that describes when a transient behavior is likely to appear, predicts what factors enhance long transients, and describes the characteristics of this transient behavior. A theory of long ecological transients is a counterpart to the related question of tipping points, where previous work based on an analysis of simple bifurcations has provided broad insights.

ADVANCES: Just as ideas based on the saddle-node bifurcation provide a basis for understanding tipping points, a suite of ideas from dynamical systems provides a way to organize a systematic study of transient dynamics in ecological systems. As illustrated in the figure, a relatively small number of ideas from dynamical systems are used to categorize the different ways that transients can arise. Translating these abstract results from dynamical systems into observations about both ecological models and ecological system dynamics, it is possible to understand when transients are likely to occur and the various properties of these transients, with implications for ecosystem management and basic ecological theory. Transients can provide an explanation for observed regime shifts that does not depend on underlying environmental changes. Systems that continually change rapidly between different long-lasting dynamics, such as insect outbreaks, may most usefully be viewed using the framework of long transients.

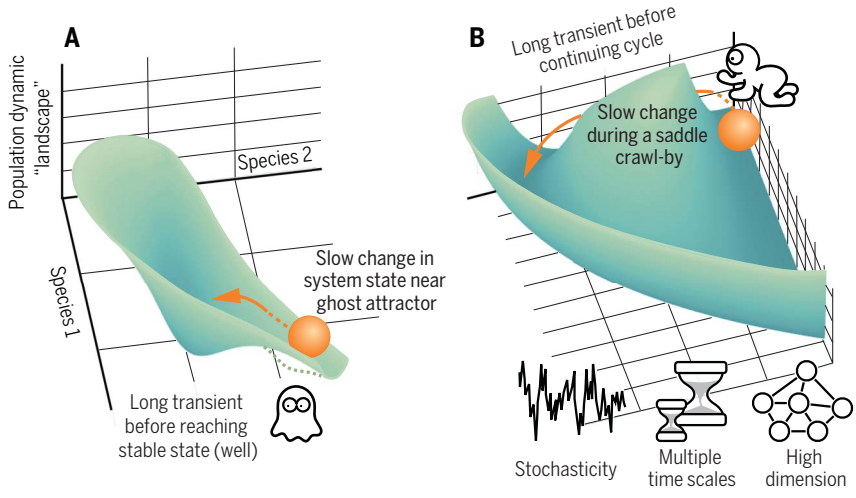
ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aat6412

An initial focus on conceptual systems, such as two-species systems, establishes the ubiquity of transients and an understanding of what ecological aspects can lead to transients, including the presence of multiple time scales and particular nonlinear interactions. The influences of stochasticity and more realistic higher-dimensional dynamics are shown to increase the likelihood, and possibly the temporal extent, of transient dynamics.

OUTLOOK: The development of such a framework for organizing the study of transients in ecological systems opens up a number of avenues for future research and application. The approach we describe also raises important questions for further development in dynamical systems. We have not, for example, emphasized nonautonomous systems, which may be required to understand the implications of a changing environment for transients. Systems with explicit time dependence as well as stochastic nonlinear systems still present great mathematical challenges.

Implications for management and basic ecological understanding depend on both the results we describe and future developments. A recognition of the difficulty of prediction caused by long transients, and of the corresponding need to match dynamics to transient behaviors of models, shows that basing either management or interpretation of ecological observations only on long-term dynamics can be seriously flawed. ■



Two ways that long transients arise in ecology, illustrated as a ball rolling downhill. (**A**) Slow transition away from a ghost attractor: a state that is not an equilibrium, but would be under slightly different conditions. (**B**) Lingering near a saddle: a state that is attracting from some directions but repelling from others. Additional factors such as stochasticity, multiple time scales, and high system dimension can extend transients.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: amhastings@ucdavis.edu
Cite this article as A. Hastings *et al.*, *Science* **361**, eaat6412 (2018). DOI: 10.1126/science.aat6412

REVIEW

ECOLOGY

Transient phenomena in ecology

Alan Hastings[1]*, Karen C. Abbott[2], Kim Cuddington[3], Tessa Francis[4], Gabriel Gellner[5], Ying-Cheng Lai[6], Andrew Morozov[7,8], Sergei Petrovskii[7], Katherine Scranton[9], Mary Lou Zeeman[10]

The importance of transient dynamics in ecological systems and in the models that describe them has become increasingly recognized. However, previous work has typically treated each instance of these dynamics separately. We review both empirical examples and model systems, and outline a classification of transient dynamics based on ideas and concepts from dynamical systems theory. This classification provides ways to understand the likelihood of transients for particular systems, and to guide investigations to determine the timing of sudden switches in dynamics and other characteristics of transients. Implications for both management and underlying ecological theories emerge.

Understanding ecological dynamics over relevant time scales underpins almost all major questions in ecology, such as explanations for observed distributions and abundances of species, population changes through time, and management of ecological systems. There is a growing recognition that dynamics on ecological time scales, called transients, may be different from asymptotic dynamics. The inherent impermanence of transients means that an ecological system in a transient state can change abruptly, even without any underlying change in environmental conditions (parameters). Conversely, the possibility of long transients implies that an ecological system may remain far from its asymptotic behavior for a long time. Insect outbreaks (*1*) such as that of the spruce budworm (*2*), where dynamics shift markedly over relatively short time scales, provide an important class of examples.

Thus, understanding the implications of transients for ecology depends on understanding potential rapid transitions between two kinds of dynamics, the behavior of systems far from their final dynamics, and the underlying time scales for these transitions. However, with the current lack of a systematic framework to facilitate understanding of transient dynamics, each example appears novel and idiosyncratic. Concepts from dynamical systems (Table 1) can provide tools for a more systematic approach to the incorporation of transient dynamics in ecological models and theories, as well as guide applications to natural and managed systems. Tools will emerge for understanding which ecological factors produce long transients, and for identifying appropriate responses to the possibility of sudden system changes in management and in experimental and observational studies.

A major ecological question is how to relate observations of changes in dynamics to underlying causes. With transients there may be no underlying proximal cause of a sudden change in dynamics. There may have been no underlying environmental change, or the change may have occurred quite far in the past. In contrast, identification of the proximal factors responsible for regime shifts has been a major focus of attention over the past two decades (*3*, *4*). Regime shifts may occur as a result of slow, directional change in ecological parameters, especially when such a change leads to a "bifurcation" of the ecosystem properties (e.g., a disappearance of a stable steady state) (*3*, *4*), also known as a "tipping point." In turn, the directional change in parameter values is often assumed to result from an exogenous process such as, for instance, global climate change. Intense study of one kind of exogenously triggered regime shift (those caused by saddle-node bifurcations) has provided important insights (*3*, *5*, *6*) across a range of ecological systems. There is, however, a growing body of evidence that we review here, from both empirical and modeling studies, suggesting alternative underlying mechanisms for some regime shifts.

The approach for understanding regime shifts can be extended to a much broader range of phenomena and systems by focusing on transients in ecological systems, where once again ideas from dynamical systems can organize what may at first appear to be a disparate set of observations and explanations. In the cases we focus on here, the ecological dynamics are essentially transient (*7*–*12*) and shifts occur in the absence of any clear trend in the environmental properties. Ecological transients can arise for a number of reasons, including responses to environmental fluctuation as well as a variety of human interventions. Some transients are short; others can last for a very long time. An ecosystem exhibiting long transient behavior would typically show an apparently stable dynamic (e.g., periodic oscillations, as in Fig. 1, A, D, and E) over time that may span dozens or even hundreds of generations before experiencing a sudden transition to another state (e.g., extinction) or another regime (e.g., oscillations with a very different mean value). Therefore, long transients may provide

[1]Department of Environmental Science and Policy, University of California, Davis, CA 95616, USA. [2]Department of Biology, Case Western Reserve University, Cleveland, OH 44106, USA. [3]Department of Biology, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. [4]Puget Sound Institute, University of Washington, Tacoma, WA 98421, USA. [5]Department of Biology, Colorado State University, Fort Collins, CO 80523, USA. [6]School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85287, USA. [7]Department of Mathematics, University of Leicester, Leicester LE1 7RH, UK. [8]Shirshov Institute of Oceanology, Moscow 117851, Russia. [9]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095, USA. [10]Department of Mathematics, Bowdoin College, Brunswick, ME 04011, USA.
***Corresponding author. Email: amhastings@ucdavis.edu**

Table 1. Key concepts used in this paper. See (*70*, *71*) for further elaboration of ideas from dynamical systems.

Term	Definition
Asymptotic dynamics	The behavior that a system will eventually exhibit and then retain indefinitely if unperturbed (i.e., dynamics that are not transient). Examples would include equilibria or limit cycles of predator-prey systems.
Bifurcation	A qualitative change in a system's asymptotic dynamics as a parameter is varied, caused by gain, loss, or change in stability of an invariant set. Examples are crises, Hopf bifurcations, and saddle-node bifurcations.
Regime shift	A qualitative change in a system's dynamics after a long period of apparent stasis. Can occur at tipping points where a bifurcation is crossed, or at a transition from transient dynamics to asymptotic dynamics (or from one transient to another).
Tipping point	The conditions (or value of a changing parameter) at which a bifurcation occurs, producing qualitatively different asymptotic behavior.
Transient	Nonasymptotic dynamics.
Long transient	Nonasymptotic dynamics that persist over ecologically relevant time scales of roughly dozens of generations (or longer).

an alternative explanation of ecological regime shifts.

Transients are not an isolated phenomenon but are related to other aspects of the dynamics of ecological systems that provide challenges for long-term prediction. With transient dynamics, the difficulty of predicting the timing of the shift between dynamic behaviors is compounded by the difficulty of determining asymptotic behavior from observations of short-term behavior (or the converse). Chaotic dynamics limit the time over which accurate predictions are possible (*13*–*15*). The permanent influence of external and internal noise on population dynamics also substantially reduces ecological predictability in a number of ways (*16*, *17*). Ecological predictions are further complicated by regime shifts (*5*, *6*) that occur as underlying environmental conditions slowly change. As a result, any conclusions or estimates made on the basis of observations preceding the regime shift simply become irrelevant after the shift. Regime shifts can often result in catastrophic changes in the ecosystem structure and function, in particular leading to species extinction and biodiversity loss.

Although long transients are often observed in ecological data (Table 2) and have been seen in many different models in ecology (*7*, *18*, *19*) as well as in neuroscience (*20*, *21*) and other natural sciences (*22*), a systematic consideration of this highly relevant phenomenon has been missing so far. Additionally, there has been some confusion about the relationship between regime shifts and long transients. We begin with an overview of ideas from dynamical systems that show why transients are a universal feature of ecological systems. We propose a simple classification scheme that shows that the mechanisms producing transients can be put into a small number of classes. This classification thus provides a new unified framework for incorporating transients into interpretations of ecological dynamics as well as into management responses. We emphasize that we can view a system as moving between transients, especially if we change the time scale we are focusing on. Additionally, we provide a road map for future investigations based on open challenges in the study of transient dynamics.

Classification and mechanisms

The unifying principle underlying past studies of long transients is a focus on multiple time scales (*23*). One example is regime shifts where slow parameter changes eventually lead to relatively rapid shifts in the state of an ecosystem. We extend this view in two critically important ways: We ask about the dynamics on both of these time scales, and we include other ways in which transients arise. Beyond this emphasis on multiple scales, we emphasize that the ecologically relevant time scales are typically not the asymptotic time scales that have been the focus of many ecological modeling studies and that form the basis of theory on which many empirical studies rest. Nor are very short time scales the appropriate focus.

The tools of dynamical systems provide the means for a systematic approach to long transients. Thus, we first review concepts from dynamical systems (*23*) that underlie the more traditional view of ecological systems representing and being represented by the asymptotic behavior of mathematical models. The simplest long-term, or asymptotic, behavior would be a stable equilibrium; a slightly more complex possibility would be a stable limit cycle. A cycle or an equilibrium are both examples of invariant sets: If the system is at an equilibrium or on a cycle, it will remain there forever in the absence of any perturbation or change in parameter values. There are also more complex invariant sets, including chaotic ones. Under the traditional view of ecological systems, on time scales that are relevant for understanding these systems, the focus should not only be on invariant sets; it should also be limited to stable invariant sets that are approached through time. A major limitation of this view is that the relevant time scale for important ecological questions may be short enough that the asymptotic behavior is not an appropriate description. We can, however, still use ideas from dynamical systems to understand and classify the behavior of ecological systems on these shorter (but not very short) time scales.

There is a broad range of transient patterns in real ecological systems, likely caused by a range of mechanisms (Table 2). We can classify these mechanisms into two general categories: those that occur in the vicinity of an invariant set, and those that do not. Within this broad classification we also can identify properties that make

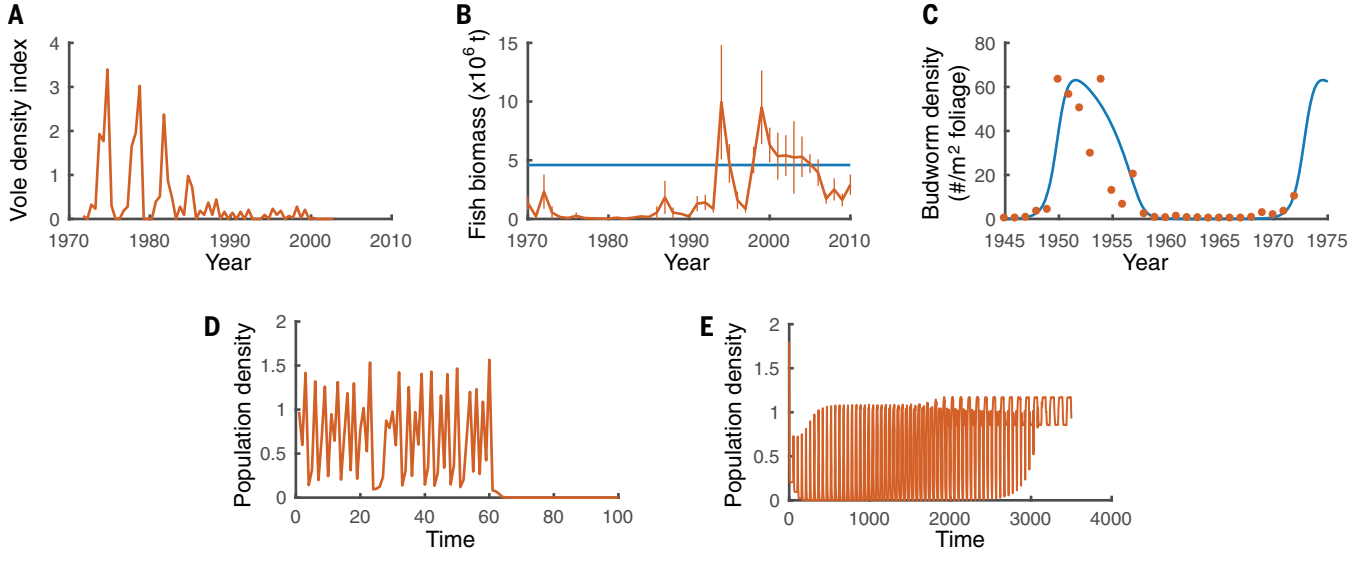


Fig. 1. Examples of transient dynamics. (**A** to **C**) Empirical examples of regime shifts occurring after long transient dynamics. (A) Population abundance of voles in northern Sweden, showing a transition from large-amplitude periodic oscillations to nearly steady-state dynamics [redrawn from (*67*)]. (B) Biomass of forage fishes in the eastern Scotian Shelf ecosystem; a low-density steady state changes to a dynamical regime with a much higher average density [blue line is the estimated carrying capacity; error bars are SEM; redrawn from (*27*)]. (C) Spruce budworm [dots; data from (*68*)] has a much faster generation time than its host tree, resulting in extended periods of low budworm density interrupted by outbreaks. A model (blue line) with fast budworm dynamics and slow foliage dynamics shows qualitative agreement with the data (*2*). (**D** and **E**) Examples of long transients on population dynamics models: (D) apparently sustainable chaotic oscillation suddenly results in species extinction (*18*); (E) large-amplitude periodic oscillations that persist over hundreds of generations suddenly transition to oscillations with a much smaller amplitude and a very different mean (*19*).

Table 2. Empirical evidence for long ecological transients.

Population(s)	Observed pattern	Duration	
		Generations	Years
Laboratory population of beetles (*Tribolium* spp.) (*25*)	Switch from a regime with an almost constant density to large-amplitude oscillations	15	~1.5 (70 weeks)
Growth of macrophytes in shallow eutrophic lakes in the Netherlands (*46*)	Switch from a macrophyte-dominated state to a turbid water state	1 to 5	1 to 5
Population of large-bodied benthic fishes on the Scotian Shelf of Canada's east coast (*27*)	Switch from a forage fish (and macroinvertebrate)–dominated state to a benthic fish–dominated state	5 to 8	20
Coral and microalgae in the Caribbean (*47*, *48*)	Shifts from coral to macroalgal dominance on coral reefs	20 to 25 (corals); 50 to 100 (macroalgae)	10
Voles, grouse in Europe (*59*)	Switch between cyclic and noncyclic regimes, or between cyclic regimes with different periodicity	60 (voles); 20 to 30 (lemmings); 5 (grouse)	~30
Dungeness crab (*Cancer magister*) (*53*)	Large-amplitude transient oscillations with further relaxation to equilibrium	10 to 15	45
Zooplankton-algal interactions in temperate lakes in Germany (*26*)	Variation of amplitude and period of predator-prey oscillations across the season	80 to 100 (algae); 5 to 8 (zooplankton)	1
Planktonic species in chemostat and temperate lakes (*72*)	Long-term variation of species densities, with extinction of some species	40 to 100	~0.05 to 0.15 (3 to 8 weeks)
Laboratory microbial communities (*56*)	Slow switch between alternative stable states	20 to 40	0.11 to 0.21 (6 to 12 weeks)
Grass community in abandoned agricultural fields in the Netherlands (*57*)	Long-term existence of a large number of alternative transient states	10	9
Extinction debt phenomena as a consequence of habitat loss [plants, birds, fish, lichens, and others (*60*)]	Long-term extinction of populations, occurring either steadily or via oscillations	20 to 100 (or more)	1 to 100
Fish and invertebrates in watersheds in western North Carolina, USA (*49*)	Influence of past habitat structure on present biodiversity patterns after restoration	10 to 20 (fish); 40 (invertebrates)	40
Modeled spruce budworm outbreaks in balsam fir forests (*2*)	Budworm outbreaks driven by slow changes in condition of fir foliage	5 (refoliation); 50+ (budworm)	50

a system particularly prone to long transients, such as the presence of multiple time scales, high dimensionality, and stochasticity.

We call a dynamical regime (e.g., a nearly constant state or persistent cycles) a long transient if it exhibits the following two properties:

1) The dynamical regime persists for a sufficiently long time and is quasi-stable (approached over shorter time scales), rather than actually stable. Thus, if the dynamics are observed for a sufficiently long time (in appropriate time units, e.g., generations of a relevant species), a clearly seen transition eventually occurs to another equilibrium or dynamic regime.

2) The transition between the regimes occurs on a time scale much shorter than the time of existence of the quasi-stable regime. In other words, the dynamics both before and after the transition last much longer than the time of transition.

Below, we consider a few simple models that exhibit long transients with somewhat different properties. We use these to formalize our classification and describe the mechanisms underlying the long transients.

Ghosts and crawl-bys

One class of long transients arises when a system is near a bifurcation. If we imagine a system's dynamics represented as a ball rolling on an uneven surface, wells correspond to stable equilibria and peaks to unstable ones. If placed into a well, the ball will roll to a stable equilibrium (Fig. 2A). Consider now the situation where the surface is

being gradually deformed in such a way that one of the wells becomes more and more shallow. Eventually the system passes the tipping point at which the stable equilibrium at the bottom of this well and the adjacent unstable equilibrium both disappear, and the ball starts rolling down the slope (Fig. 2B). However, how fast it starts moving away—or, in other words, how much time it remains in the vicinity of the location where the stable equilibrium was before the bifurcation—depends on the flatness of the slope. The flatter the surface is, the longer the ball stays close to its original location before moving away: The long transient emerges. Although beyond the tipping point the system does not possess an equilibrium at this long-lasting state, for a considerable time its dynamics mimic the dynamics of the system with an attractor here (Fig. 2, C and D). We call this situation a ghost attractor (*24*) or simply a ghost.

The origin of a ghost attractor and an example of the long transients it can cause are shown in Fig. 2, A to D. To understand the importance of this effect, imagine, for instance, that competitor 1 in Fig. 2, B and D, is a native species competing with an invader. At the early stages of invasion, we expect the native species to be abundant and the invader rare. From these initial conditions, the system can spend considerable time in this state, even if the ultimate asymptotic result is that the invader excludes the native species (as in Fig. 2B). This occurs because the invaded system has conditions that are close to, but distinct from, those that would have allowed the invader and the native species to coexist (Fig. 2, A and C). Correspondingly, if the system is only monitored on an intermediate time scale, this long transient dynamic may give an impression that both species will coexist indefinitely—a conclusion that would obviously be erroneous on a longer time scale (Fig. 2D).

The long transient dynamics in Fig. 2 occur because of a bifurcation that results in the disappearance of a stable equilibrium. Beyond the bifurcation, there is no longer an equilibrium in the vicinity of what is now a ghost, but the system may still spend a long time in this vicinity. In other words, the long transient occurs without an invariant set. In contrast, the second class of transients we define requires the existence of an unstable equilibrium (more specifically, the existence of a saddle). The system approaches the saddle along a stable direction and spends a long time near the saddle. We call this transient a crawl-by.

We find examples of this type of long transient in predator-prey systems (Fig. 3). Note that dynamics with similar properties are observed in more realistic and more complicated models (*25*, *26*) and are corroborated by some field and laboratory data (*25*–*27*) (see also Fig. 1B), which points at the generality of the suggested mechanism.

Note that crawl-bys and ghosts appear to be similar: Having spent some considerable time in the vicinity of its original position, the system (e.g., the ball) eventually moves away. However,

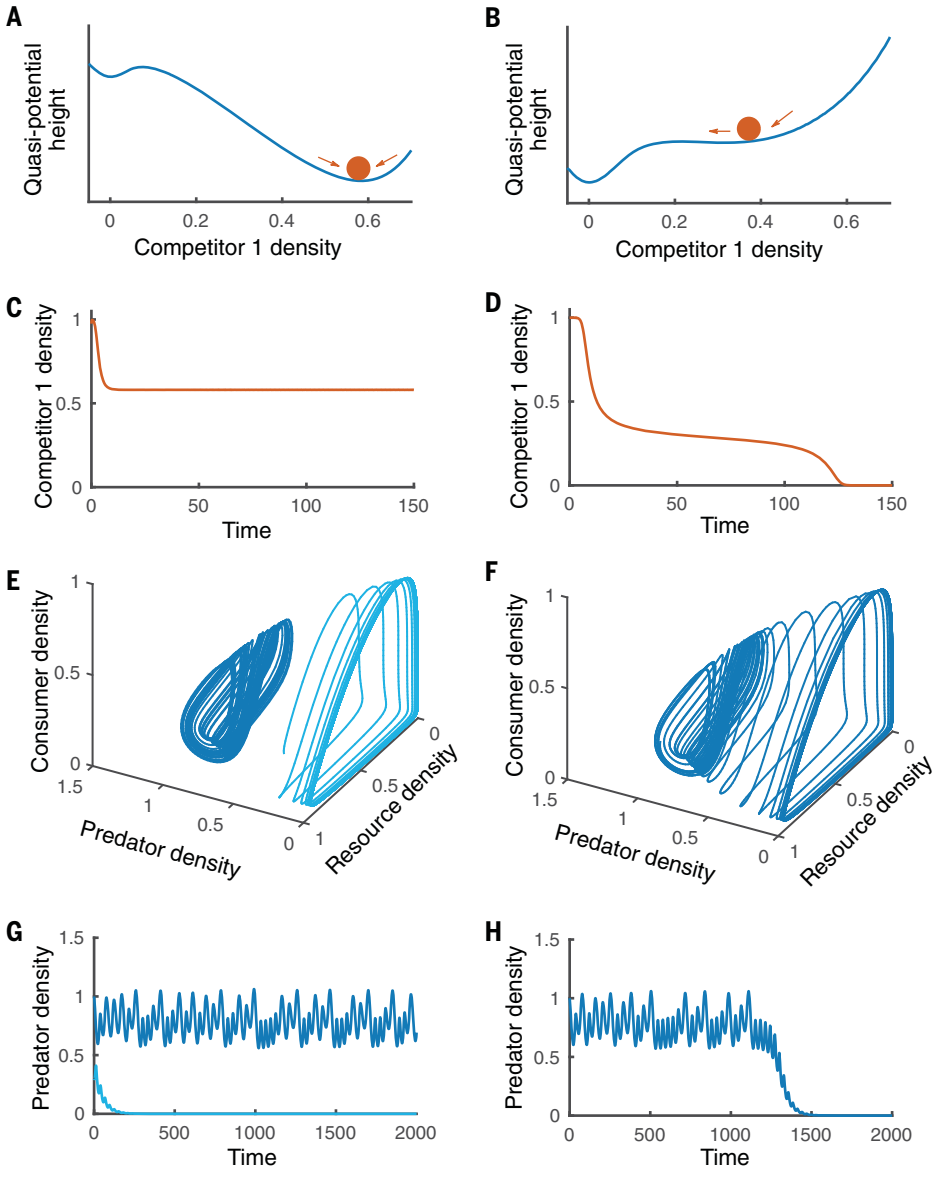


Fig. 2. Ghost attractors. Illustration of ghost attractors in a two-species competition model (**A** to **D**) and a resource-consumer-predator model (**E** to **H**). In the left column [(A), (C), (E), and (G)], there are two stable invariant sets and no ghost attractors. In the right column [(B), (D), (F), and (H)], there is a single stable invariant set, plus a ghost attractor that causes long transients. A bifurcation (tipping point) occurs for parameter values intermediate to these two cases; at this bifurcation, one stable state is lost and a ghost attractor takes its place. [(A) and (B)] Dynamics of one of the competitors depicted as a ball on a quasi-potential surface. In (A), a ball to the right of the hump at 0.07 will tend to roll toward the stable equilibrium (well) at 0.58, as in time series (C). In (B), the surface is relatively flat, rather than containing a well, to the right of ~0.1; a ball to the right will eventually roll to the stable equilibrium at 0 but will roll very slowly on the flat part of the surface, generating a long transient. There is a ghost attractor at a density around 0.3, which is visible in time series (D). [(E) to (H)] The same phenomenon with more complex invariant sets: [(E) and (G)] The system shows bistability where a chaotic three-species attractor (dark blue) coexists with a stable consumer-resource limit cycle with no predators (light blue); dark and light trajectories differ only in their initial conditions. [(F) and (H)] For parameter values on the other side of a bifurcation that turns the chaotic attractor into a chaotic saddle, any trajectory will eventually converge to the stable limit cycle, which is now the global attractor. However, convergence can be slow, as seen in (H), because the chaotic set is now a ghost. Models are as follows: [(A) to (D)] Competitor 1 is v and competes with species u: $du/dt = u(1-u) - a_{12}u^n v$, $dv/dt = \gamma[v(1-v) - a_{21}u^n v]$ with $a_{12} = 0.9$, $a_{21} = 1.1$, $\gamma = 10$, and $n = 3$ [(A) and (C)] or $n = 1.55$ [(B) and (D)]; [(E) to (H)] from (*28*, *29*), where the resource is R, consumer C, and predator P: $dR/dt = R[1 - (R/K)] - x_c y_c CR/(R + R_0)$, $dC/dt = x_c C\{[y_c R/(R + R_0)] - 1\} - x_p y_p PC/(C + C_0)$, $dP/dt = x_p P\{[y_p C/(C + C_0)] - 1\}$ with $x_c = 0.4$, $y_c = 2.009$, $x_p = 0.08$, $y_p = 2.876$, $R_0 = 0.16129$, $C_0 = 0.5$, and $K = 0.99$ [(E) and (G)] or $K = 1$ [(F) and (H)]. Quasi-potentials in (A) and (B) were computed using (*69*).

a distinction appears when the history of the system is taken into account. For a system to be influenced by a ghost, its initial state must be near the ghost (as in Fig. 2B) or more extreme, such that it passes by the ghost en route to another state (as in Fig. 2D). One reason a system's history might place it near a ghost is if that system recently underwent a change in conditions that caused the ghost attractor to appear. Individual crawl-bys can also occur if the history of the system places it on track to closely approach a saddle, but crawl-bys may also repeat in perpetuity, as in Fig. 2, C and D. This occurs because the saddles that give rise to crawl-bys are always attracting from some directions, whereas ghosts may or may not have attracting directions.

The mechanisms described above that cause long transients are not restricted to simple dynamics such as steady states or limit cycles. Similar effects can be seen in cases of chaotic dynamics. An illuminating example is given by a resource-consumer-predator system (*28*, *29*). In a certain parameter range, this system exhibits chaotic dynamics [see the chaotic attractor (dark blue) in Fig. 2, E and G]. However, a change in parameter values (e.g., an increase in the resource species' carrying capacity) can bring the system to a bifurcation at which the strange attractor disappears (*29*). Beyond this tipping point, the chaotic dynamic is not sustainable any more; it becomes transient and eventually converges to a periodic oscillation with a stable limit cycle (Fig. 2, F and H). However, this convergence is slow, so that the dynamics remain essentially chaotic over a long time. Similar dynamics are observed in time-discrete systems (*20*). This behavior is apparently similar to the crawl-by near a saddle point, and indeed the term "chaotic saddle" is used in the physics literature to refer to a nonattracting dynamical invariant set responsible for transient chaos (*30*, *31*) (in dynamical systems theory, it is also called a chaotic supertransient). More generally, a common dynamical mechanism for transient chaos is crisis (*30*), a type of global bifurcation that changes the nature of the underlying chaotic invariant set.

An important point is that one property common to long transients caused by ghost attractors and chaotic crawl-bys is that the system is just beyond the tipping point. Thus, if a parameter controlling a system has pushed the system just past a tipping point, there may not be a sudden change; instead, a long transient may result.

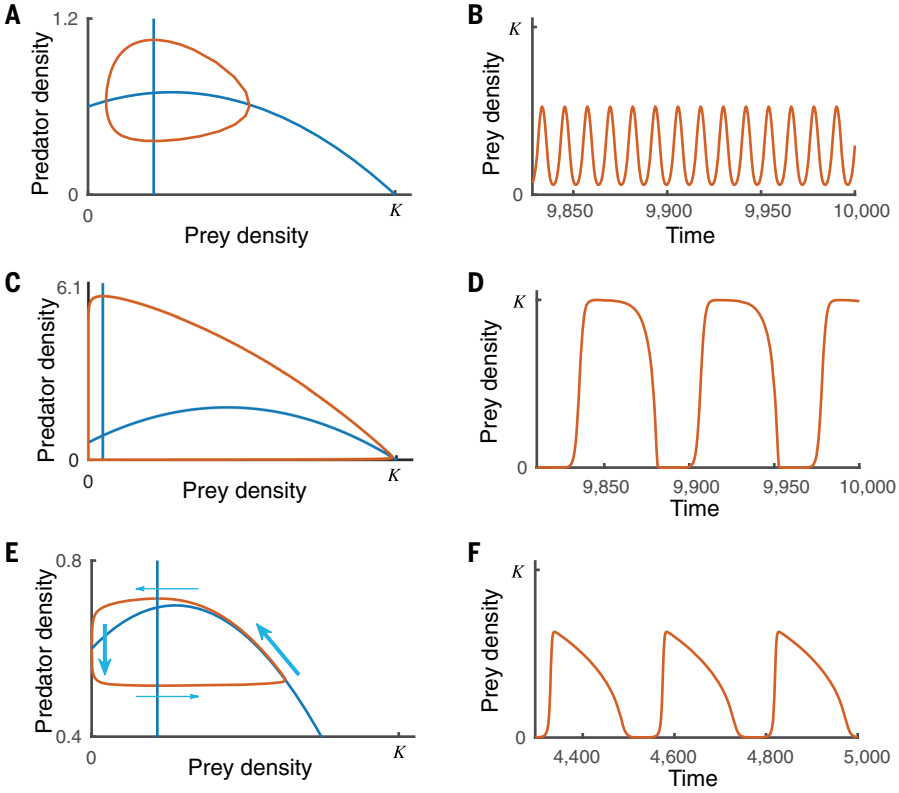


Fig. 3. Predator-prey dynamics with and without transients. Predator-prey dynamics without long transients (**A** and **B**), with long transients due to crawl-bys (**C** and **D**), and with long transients due to slow-fast dynamics (**E** and **F**). In (A), (C), and (E), the intersection of the predator's and prey's isoclines (blue lines) produces a coexistence equilibrium. When the prey's predator-free carrying capacity K is beyond a threshold (Hopf bifurcation), the system exhibits limit cycles around this equilibrium. [(A) and (B)] For K just beyond this threshold, relatively small limit cycles occur and there are no long transients. [(C) and (D)] With an increase in K, the cycle grows in size and closely approaches the two saddle points at (0,0) and (K,0). In (D), crawl-bys are visible at 0 and K. (E) When predator (slow) and prey (fast) dynamics occur on very different time scales, the shape of the cycle changes, and more horizontal parts of the cycle (thin arrows) proceed much more quickly than more vertical parts (thick arrows). (F) The corresponding time series for the prey shows long transients at 0 and higher prey density. The difference between (F) and (B) is entirely due to the slower predator dynamics in (F). In all panels, for prey species N and predator P, $dN/dt = \alpha N[1 - (N/K)] - \gamma NP/(N + H)$, $dP/dt = \varepsilon\{[\nu\gamma NP/(N + h)] - mP\}$ with $\gamma = 2.5$, $h = 1$, $\nu = 0.5$, $m = 0.4$. In (A), (B), (E), and (F), $\alpha = 1.5$, $K = 2.2$; in (C), $\alpha = 1.5$, $K = 10$; in (D), $\alpha = 0.8$, $K = 15$; in (A) to (D), $\varepsilon = 1$; in (E) and (F), $\varepsilon = 0.01$.

Slow-fast systems

Much of the literature on tipping points considers multiple time scales: fast intrinsic dynamics affected by a slow-changing external factor. However, some systems have multiple time scales within their intrinsic dynamics. This can also lead to transients, as in a prey-predator system written in its general form:

$$\frac{dN(t)}{dt} = f(N, P, \varepsilon)$$

$$\frac{dP(t)}{dt} = \varepsilon g(N, P, \varepsilon) \quad (1)$$

where $\varepsilon << 1$ is a non-negative dimensionless parameter that quantifies the difference between orders of magnitude for the time scales of prey (N) and predator (P) (*32*), and f and g are the growth rate of the prey population on the natural time scale for the prey and the growth rate of the predator population on the natural time scale for the predator, respectively. Such a difference is common in resource-consumer interactions. For example, univoltine insect herbivores that feed on trees have much faster population dynamics than their hosts. Reproduction and mortality rates of zooplankton are typically lower by one to two orders of magnitude than the corresponding rates of phytoplankton on which the zooplankton feed. Similar differences exist for birds and insects, foxes and voles, etc. (*33*).

Viewed on the slow time scale, the prey population evolves quickly and is always at its equilibrium, with the predator population acting essentially as a slowly changing parameter with dynamics determined by the predator equation. The net result is alternation between long periods of relative stasis and periods of rapid change. An almost steady-state dynamic of prey at a very low density accompanied by a gradual decrease in the predator density (as shown by the left side of the cycle in Fig. 3E and each trough in Fig. 3F) can go on for hundreds of generations of prey before suddenly changing to an outbreak in the prey population. The next phase is a slow, gradual decrease in the prey population along with a slow increase in the predator population (the right side of the cycle, and peaks in the prey time series) before accelerating to a fast drop in the prey density. The difference between this dynamic and the transients described above, in which there is only

one intrinsic time scale, is whether the slowly changing variable is viewed as internal or external to the system. This is important because slowly changing variables are often considered the result of human actions or a changing environment; they could alternatively be viewed as part of the system itself. These systems with inherent multiple time scales lead to the view that whether we think of a system as having transients may depend on the time scale of ecological interest relative to the time scales embodied in the system.

Transients in high-dimensional systems

Most systems outside a laboratory or experimental environment are very high-dimensional because of the presence of space or time delays that greatly increase the potential for transients, including very long transients. In the examples considered so far, all the processes or forces shaping the dynamics are instantaneous and local in space. In real-world systems, it is not always so.

Time delays are a common property in ecological dynamics resulting from processes and mechanisms such as nutrient recycling (*34*, *35*), maternal effects (*36*, *37*), or development in stage-structured populations. Time delays were shown to lead to the emergence of long transients in a few modeling studies (*19*, *38*), and there is a certain similarity between delay-caused long transients and those caused by ghost attractors. Systems with time delay are different from instantaneous systems not only because of different processes taken into account, but also from the viewpoint of dynamical systems theory as the phase-space argument and the corresponding analysis become irrelevant. In a general case, even a low-dimensional (e.g., two-species) system with delay is equivalent to an infinite-dimensional instantaneous system (*39*, *40*). These findings suggest that time delay is a separate mechanism that can result in long transient dynamics.

Spatiotemporal dynamical systems are necessarily high-dimensional, and the transient time can greatly increase with the system size. An early study (*7*) reported extremely long transients in such systems (Fig. 4A). In systems described by a coupled map lattice, the transient time can increase exponentially with the system size or faster (*41*), leading to supertransients (*42*).

Effect of noise

Until now, we have considered long transient dynamics in deterministic settings, absent noise or stochasticity. In natural systems, noise and random disturbances are inevitable and may create or extend transients. In other cases, stochasticity may essentially eliminate transient dynamics; as we have emphasized, the practical impact of stochasticity once again will depend on the time scale of ecological interest relative to the time scales of the system dynamics.

Noise may affect a system with existing long transient dynamics caused, for example, by a ghost or crawl-by. In transient dynamics caused by a crawl-by, such as the limit cycles of a predator-prey system (Fig. 3, C and D), small populations very close to the saddle at (0,0) are vulnerable to stochastic extinctions, where random events may move the system to one of the saddle points, causing either the prey or predator population (or both) to go extinct. Stochasticity in the system near the saddle also has the potential to alter the length of the transient period, widening the distribution of resulting durations of the transient period or times to convergence (*43*). Stochasticity in the system near a ghost attractor also widens the distribution of transient periods, depending on the steepness of the surface around the ghost. Noise that is skewed "uphill" will lengthen the transient, dooming the ball to repeatedly roll nearer the ghost (Fig. 2B). Noise can also induce sustained transients or oscillations in a system that would exhibit damped oscillations to an equilibrium in the absence of noise (*44*, *45*) (Fig. 4B). Noise can also provide a mechanism for transient dynamics of a system to become long-lived (Fig. 4B). For systems with transient chaos, the interaction with stochasticity can be even more complex (*42*).

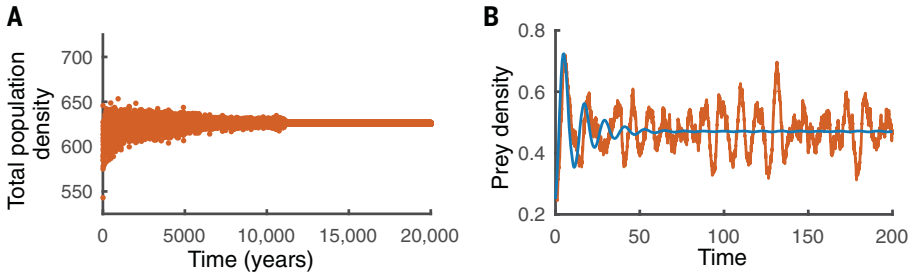


Fig. 4. Examples of additional mechanisms leading to long transients. (**A**) Spatial structure in a simple population model leads to very long transients when the local population growth rate is high [from (*7*); local dynamics are governed by $N_{t+1} = N_t \exp[r(1 - N_t)]$ with $r = 3.5$; the total population density summed across all localities is plotted here]. (**B**) For these parameter values ($\alpha = 1.5$, $K = 1.5$, $\gamma = 2.5$, $h = 1$, $\nu = 0.5$, $m = 0.4$, $\varepsilon = 1$), the deterministic predator-prey model from Fig. 3 exhibits short transient cycles, then converges to a stable equilibrium point (blue curve). However, when stochasticity is added, the same model will exhibit sustained cycles with approximately the same period (red line; here, stochasticity was incorporated by representing the prey's intrinsic growth rate, α, as a random variable with mean 1.5).

Transients in the real world

The systematic classification of long transient types and mechanisms conducted here provides a framework for recognizing and understanding these dynamics in observed natural systems (Tables 1 to 3). Note that our classification does not include nonautonomous systems, not because long transients do not occur in nonautonomous systems, but because their classification and discussion warrants treatment beyond the scope of the present review. In this section we describe how empirically observed behavior may be the result of long transients in a wide variety of situations; through examples, we emphasize implications for management.

An empirical example of long transients due to a ghost attractor (similar to that presented in Fig. 2) is the well-documented switch from a macrophyte-dominated state to a turbid water state in freshwater lakes in the Netherlands (*46*). The study tracked about 70 shallow lakes after a water drawdown that stimulated macrophyte growth, temporarily creating a macrophyte-dominated, clear water attracting state. When water levels subsequently rose, some of the lakes returned immediately to a turbid state, while others lingered for more than 4 years in the clear water state that was no longer stable. In other words, the physical modification to the system caused by the changes in water level resulted in the formation of a clear water ghost attractor that slowed movement toward the turbid water attractor, sometimes quite substantially. A similar mechanism of long transients due to ghost attractors may underlie the transition from coral to macroalgal dominance reported for Caribbean coral reefs (*47*, *48*), the shift from a forage fish state to a state dominated by large-bodied benthic fish species in the Scotian Shelf of Canada's east coast (*27*), and the shifts between populations of fish and invertebrates in watersheds in western North Carolina after habitat restoration (*49*).

Long transients due to crawling past a saddle are often observed in planktonic ecosystems—in particular, in the interactions between phytoplankton and zooplankton—creating oscillations in which periods of high population density alternate with long periods of low density (*26*). Other examples of crawl-bys are given by patterns of cyclic succession reported in a number of ecosystems, including competition in communities of side-blotched lizards (*50*), coral reef invertebrates (*51*), and heather-moss-bearberry succession (*52*). In each of these systems, a long dominance of one species is observed before its displacement by the next competitor in the cycle.

Empirical examples of long transients related to slow-fast systems include a number of observations of univoltine insect herbivores that feed on trees (*2*) (Fig. 1C). At short time scales of a few insect generations, the tree density is approximately constant. However, on longer time scales, the impact of the growing insect population may become high enough to cause a sudden collapse in the quantity or quality of foliage, resulting in a regime shift.

Real ecosystems are often disturbed by noise that can trigger patterns of long transients. A notable example includes the population dynamics

Table 3. Overview of long transient (LT) classification and mechanisms.

Type of LT	Relationship to invariant set	Relationship to bifurcation	Dynamics mimicked by LT	Possibility of recurrent LTs?	Biological example
Ghost (Fig. 2)	No invariant set	Occurs past a bifurcation where stable equilibrium is lost	Equilibrium, cycles, chaos	No	Forage fish (*27*) (Fig. 3B)
Crawl-by (Fig. 3, C and D)	Caused by saddle-type invariant set	None necessary	Equilibrium, cycles, chaos	Yes	Phytoplankton-grazer models (*26*)
Slow-fast systems (Fig. 3, E and F)	None necessary	Multiple time scales	Periodic or aperiodic cycles	Yes, if invariant set(s) present	Univoltine insects (*2*) (Fig. 3C)
High dimension (e.g., time delays, space) (Fig. 4A)	None necessary	None necessary	Equilibrium, cycles, chaos	Yes	Chemostat microbial communities (*57*)
Stochasticity (Fig. 4B)	If invariant set present:	None necessary	Aperiodic cycles, chaos	Yes	*Cancer* crabs (*53*)
	If invariant set absent:	Past a bifurcation where cycles/chaos are lost	Quasi-periodic cycles		

of Dungeness crab, *Cancer magister*, in eight West Coast ports of the United States (*53*). By combining data analysis with modeling fitted to data, large-amplitude transient oscillations followed by relaxation to an equilibrium were shown to occur as a result of stochastic perturbations of a deterministic system with a stable state. Another example is given by an empirical study on *Tribolium* (*54*) in which random perturbations of cyclic population dynamics result in chaotic-like behavior. Seasonal dynamics create a particular structure of environmental stochasticity. For example, in plankton communities in temperate lakes, each cold season "resets" the initial conditions for the warm, growing season. This prevents the system from reaching equilibrium and thereby allows for high biodiversity transients to be the norm (*26*, *55*).

High-dimensional systems may be likely to possess long transients. For example, slow succession of patterns of biodiversity is found in experimental microbial communities in a chemostat (*56*). The precise mechanism of these observed long transients is still unclear because of the high complexity of systems containing dozens of interacting species and the existence of several time scales. Similarly, long-term existence of a large number of alternative transient states is seen in the restoration of agricultural fields (*57*), which is also characterized by a high degree of complexity.

Implications for management

The existence, identification, and forecasting of long transient dynamics in ecosystems have substantial implications for the management of ecosystems. Broadly speaking, management is aimed at maintaining or creating a desirable state of the ecosystem. The challenge is in predicting system behavior given dynamical regime uncertainty. If a system transition is detected, the important questions are what has caused it and how long it can be expected to last. What a study of long transients reveals is that a system may shift in ways that are not simply tracking underlying conditions, so a focus on asymptotic behavior without considering transients may give misleading answers.

In some cases, mechanistic mathematical models that are constructed from first principles, fitted to empirical data, and explored within realistic parameter ranges can help to identify whether an ecosystem is currently experiencing transient dynamics. For example, this was done to predict the long transients in the extinction debt of butterflies in the United Kingdom (*58*). In other cases, when it is difficult to distinguish whether observed dynamics are transient or at equilibrium, models of both possibilities can be developed to test the sensitivity of proposed management strategies to the model assumptions.

Incorporating considerations of transient system behavior into management also requires shifting perspectives about the relevant time scale. A fundamental issue is a mismatch between relevant ecological (transient) time scales and management time scales. Implementation of management plans where long transients are at play will require adjustments to accommodate the transient changes in dynamical regime.

Acknowledging transient system behavior affects management assumptions, practices, and interventions. In addition, detecting long transients, and incorporating risk analysis for long transients, requires the development and application of new tools to reflect this change in thinking. A full treatment of the management consequences and opportunities presented by long transients requires further attention beyond the present review.

Implications and future directions

Sudden changes in ecological dynamics through time represent both great challenges and opportunities for unraveling the forces that regulate ecosystem functioning. Much recent work in this vein has focused on the concept of regime shifts as a rapid response of dynamics to slow changes in environmental conditions (such as climate change, habitat destruction, resource exploitation, etc.). However, there are many examples of situations and systems that do not fit into this classical framework; in particular, a shift can suddenly occur in a seemingly constant environment. The existence of long transients explains how this may happen: As we have shown here, the ecosystem dynamics past a tipping point can be very slow, sometimes indistinguishable from the steady state for hundreds of generations ("ghosts"). Long transient dynamics can also be responsible for regime shifts in the absence of any associated tipping point, thus significantly broadening the regime shift paradigm. Finally, the dynamics of some systems with multiple time scales may best be viewed as a succession of transients.

The traditional approaches in ecological sciences are usually based on asymptotic dynamics. Here we have shown that this focus is often insufficient and sometimes irrelevant, and needs to be reconsidered in a systematic way. Long transients provide a new dimension in our understanding of observed changes in ecological dynamics. Although the existence of long transients has been previously acknowledged both in theoretical and empirical studies, any systematic approach to this phenomenon has been lacking. We bridge this gap by developing a simple classification of different types of long transient dynamics and linking empirical observations to simple prototypical models. As one important result of our investigation, we have arrived at the conclusion that both identifying long transients and understanding their implications (e.g., for ecosystem management) requires coupling across several ecologically relevant time scales.

Identifying from observations whether a natural ecosystem is close to an equilibrium or is experiencing long transient behavior constitutes a major challenge. Perhaps the easiest case is a situation where the population density shows a clear disappearance of periodic cycles of voles, lemmings, and grouse in Europe (*59*) (Fig. 1A), or the slow steady population decline in extinction debt phenomena (*60*). Less evident is the situation where the dynamics do not show a pronounced trend. In this case, one can compare characteristics of the observed community with those thought to represent equilibrium systems, such as a relatively undisturbed community of a similar type or the same community in the past. These ideas have been implemented to reveal an extinction debt caused by habitat fragmentation, by comparing the current species-area relation to historical records (*60*) and by verifying whether the species-area relation holds (*61*). As an alternative approach, recently developed techniques make it possible to build an ecological model directly from a time series by reconstructing model equations from data. A particularly promising new approach is based on compressive sensing using a powerful sparse optimization framework (*62*–*65*). Once a model is available, its properties can be analyzed, in particular to reveal long transients. However, we argue that any essential progress in this area is only likely to be achieved by combining various methods borrowed from data mining, stochastic modeling, and bifurcation theory. Another important aspect will be the exchange of ideas with other areas of biology such as neuroscience, where transients are considered important and have been studied with both data and models (*20*, *21*). Existing methods to identify transients in empirical data are not always reliable and can result in either overlooking the approaching regime shift or in false alarms, which can be very costly. This poses substantial new challenges for ecologists, mathematicians, and data analysts.

Acknowledging long-term transient behavior drastically affects our perception of ecological dynamics. First, sudden shifts in dynamics may occur in the absence of underlying parameter changes (i.e., in the absence of the tipping point). Second, analysis of ecological processes must be done across a few relevant time scales rather than focusing only on asymptotic behavior. In particular, one should take into account both fast and slow variables and feedbacks. Third, stochasticity may play a key role in generating long transients, in particular by bringing an ecosystem to the vicinity of an unstable equilibrium (causing a crawl-by) or a ghost. Finally, in the context of ecosystem management practices, it is well known that sometimes long transients may offer a window of response time that would not be available for systems with rapid switches between stable states (*66*). On the other hand, transients can add more uncertainty to the anticipation of regime shifts, because such a shift can occur without a noticeable change of parameters.

REFERENCES AND NOTES

1. P. Barbosa, J. C. Schultz, *Insect Outbreaks* (Academic Press, 1987).
2. D. Ludwig, D. D. Jones, C. S. Holling, Qualitative analysis of insect outbreak systems: The spruce budworm and forest. *J. Anim. Ecol.* **47**, 315–332 (1978). doi: doi:10.2307/3939
3. M. Scheffer, D. Straile, E. H. van Nes, H. Hosper, Climatic warming causes regime shifts in lake food webs. *Limnol. Oceanogr.* **46**, 1780–1783 (2001). doi: 10.4319/lo.2001.46.7.1780
4. S. R. Carpenter *et al*., Early warnings of regime shifts: A whole-ecosystem experiment. *Science* **332**, 1079–1082 (2011). doi: 10.1126/science.1203672; pmid: 21527677
5. V. Dakos *et al*., Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data. *PLOS ONE* **7**, e41010 (2012). doi: 10.1371/journal.pone.0041010; pmid: 22815897
6. C. Boettiger, A. Hastings, Quantifying limits to detection of early warning for critical transitions. *J. R. Soc. Interface* **9**, 2527–2539 (2012). doi: 10.1098/rsif.2012.0125; pmid: 22593100
7. A. Hastings, K. Higgins, Persistence of transients in spatially structured ecological models. *Science* **263**, 1133–1136 (1994). doi: 10.1126/science.263.5150.1133; pmid: 17831627
8. A. Hastings, Transient dynamics and persistence of ecological systems. *Ecol. Lett.* **4**, 215–220 (2001). doi: 10.1046/j.1461-0248.2001.00220.x
9. A. Hastings, Transients: The key to long-term ecological understanding? *Trends Ecol. Evol.* **19**, 39–45 (2004). doi: 10.1016/j.tree.2003.09.007; pmid: 16701224
10. A. Hastings, Timescales, dynamics, and ecological understanding [Robert H. MacArthur Award lecture]. *Ecology* **91**, 3471–3480 (2010). doi: 10.1890/10-0776.1; pmid: 21302817
11. R. S. Wilson *et al*., A typology of time-scale mismatches and behavioral interventions to diagnose and solve conservation problems. *Conserv. Biol.* **30**, 42–49 (2016). doi: 10.1111/cobi.12632; pmid: 26390368
12. A. Hastings, Timescales and the management of ecological systems. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 14568–14573 (2016). doi: 10.1073/pnas.1604974113; pmid: 27729535
13. R. M. May, Simple mathematical models with very complicated dynamics. *Nature* **261**, 459–467 (1976). doi: 10.1038/261459a0; pmid: 934280
14. A. Hastings, C. L. Hom, S. Ellner, P. Turchin, H. C. J. Godfray, Chaos in ecology: Is mother nature a strange attractor? *Annu. Rev. Ecol. Syst.* **24**, 1–33 (1993). doi: 10.1146/annurev.es.24.110193.000245
15. J. M. Cushing, R. F. Costantino, B. Dennis, R. Desharnais, S. M. Henson, *Chaos in Ecology: Experimental Nonlinear Dynamics* (Elsevier, 2002).
16. J. V. Greenman, T. G. Benton, The amplification of environmental noise in population models: Causes and consequences. *Am. Nat.* **161**, 225–239 (2003). doi: 10.1086/345784; pmid: 12675369
17. K. C. Abbott, B. C. Nolting, Alternative (un)stable states in a stochastic predator-prey model. *Ecol. Complex.* **32**, 181–195 (2016). doi: 10.1016/j.ecocom.2016.11.004
18. S. J. Schreiber, Allee effects, extinctions, and chaotic transients in simple population models. *Theor. Popul. Biol.* **64**, 201–209 (2003). doi: 10.1016/S0040-5809(03)00072-8; pmid: 12948681
19. A. Y. Morozov, M. Banerjee, S. V. Petrovskii, Long-term transients and complex dynamics of a stage-structured population with time delay and the Allee effect. *J. Theor. Biol.* **396**, 116–124 (2016). doi: 10.1016/j.jtbi.2016.02.016; pmid: 26921467
20. M. I. Rabinovich, P. Varona, A. I. Selverston, H. D. I. Abarbanel, Dynamical principles in neuroscience. *Rev. Mod. Phys.* **78**, 1213–1265 (2006). doi: 10.1103/RevModPhys.78.1213
21. M. Rabinovich, R. Huerta, G. Laurent, Transient dynamics for neural processing. *Science* **321**, 48–50 (2008). doi: 10.1126/science.1155564; pmid: 18599763
22. Y.-C. Lai, R. L. Winslow, Geometric properties of the chaotic saddle responsible for supertransients in spatiotemporal chaotic systems. *Phys. Rev. Lett.* **74**, 5208–5211 (1995). doi: 10.1103/PhysRevLett.74.5208; pmid: 10058710
23. C. Kuehn, *Multiple Time Scale Dynamics* (Springer, 2016).
24. E. M. Izhikevich, *Dynamical Systems in Neuroscience* (MIT Press, 2007).
25. J. Cushing, B. Dennis, R. Desharnais, R. Costantino, Moving towards unstable equilibrium: Saddle nodes in population systems. *J. Anim. Ecol.* **67**, 298–306 (1998). doi: 10.1046/j.1365-2656.1998.00194.x
26. C. G. Jäger, S. Diehl, C. Matauschek, C. A. Klausmeier, H. Stibor, Transient dynamics of pelagic producer-grazer systems in a gradient of nutrients and mixing depths. *Ecology* **89**, 1272–1286 (2008). doi: 10.1890/07-0347.1; pmid: 18543621
27. K. T. Frank, B. Petrie, J. A. Fisher, W. C. Leggett, Transient dynamics of an altered large marine ecosystem. *Nature* **477**, 86–89 (2011). doi: 10.1038/nature10285; pmid: 21796120
28. A. Hastings, T. Powell, Chaos in a three-species food chain. *Ecology* **72**, 896–903 (1991). doi: 10.2307/1940591
29. K. McCann, P. Yodzis, Nonlinear dynamics and population disappearances. *Am. Nat.* **144**, 873–879 (1994). doi: 10.1086/285714
30. C. Grebogi, E. Ott, J. Yorke, Crises, sudden changes in chaotic attractors, and transient chaos. *Physica D* **7**, 181–200 (1983). doi: 10.1016/0167-2789(83)90126-4
31. T. Kapitaniak, Y.-C. Lai, C. Grebogi, Metamorphosis of chaotic saddle. *Phys. Lett. A* **259**, 445–450 (1999). doi: 10.1016/S0375-9601(99)00479-X
32. S. Rinaldi, M. Scheffer, Geometric analysis of ecological models with slow and fast processes. *Ecosystems* **3**, 507–521 (2000). doi: 10.1007/s100210000045
33. J. H. Brown, J. F. Gillooly, A. P. Allen, V. M. Savage, G. B. West, Toward a metabolic theory of ecology. *Ecology* **85**, 1771–1789 (2004). doi: 10.1890/03-9000
34. T. Daufresne, L. O. Hedin, Plant coexistence depends on ecosystem nutrient cycles: Extension of the resource-ratio theory. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 9212–9217 (2005). doi: 10.1073/pnas.0406427102; pmid: 15964989
35. J. Pastor, R. D. Walker, S. Larsson, Delays in nutrient cycling and plant population oscillations. *Oikos* **112**, 698–705 (2006). doi: 10.1111/j.0030-1299.2006.14478.x
36. L. Ginzburg, D. Taneyhill, Population cycles of forest lepidoptera: A maternal effect hypothesis. *J. Anim. Ecol.* **63**, 79 (1994). doi: 10.2307/5585
37. P. Turchin, *Complex Population Dynamics: A Theoretical/Empirical Synthesis* (Princeton Univ. Press, 2003).
38. M. Jankovic, S. Petrovskii, Are time delays always destabilizing? revisiting the role of time delays and the allee effect. *Theor. Ecol.* **7**, 335–349 (2014). doi: 10.1007/s12080-014-0222-z
39. Y. Kuang, *Delay Differential Equations with Applications in Population Dynamics* (Academic Press, 1993).
40. H. Smith, *An Introduction to Delay Differential Equations with Applications to the Life Sciences* (Springer, 2011).
41. J. P. Crutchfield, K. Kaneko, Are attractors relevant to turbulence? *Phys. Rev. Lett.* **60**, 2715–2718 (1988). doi: 10.1103/PhysRevLett.60.2715; pmid: 10038433
42. Y.-C. Lai, T. T'el, *Transient Chaos: Complex Dynamics on Finite Time Scales* (Springer, 2011).
43. A. T. Winfree, Biological rhythms and the behavior of populations of coupled oscillators. *J. Theor. Biol.* **16**, 15–42 (1967). doi: 10.1016/0022-5193(67)90051-3; pmid: 6035757
44. R. M. Nisbet, W. Gurney, *Modelling Fluctuating Populations* (Wiley Interscience, 1982).
45. M. L. Zeeman *et al*., Resilience of socially valued properties of natural systems to repeated disturbance: A framework to support value-laden management decisions. *Nat. Resour. Model.* **31**, e12170 (2018). doi: 10.1111/nrm.12170
46. G. J. van Geest, H. Coops, M. Scheffer, E. van Nes, Long transients near the ghost of a stable state in eutrophic shallow lakes with fluctuating water levels. *Ecosystems* **10**, 37–47 (2005). doi: 10.1007/s10021-006-9000-0
47. J. F. Bruno, H. Sweatman, W. F. Precht, E. R. Selig, V. G. W. Schutte, Assessing evidence of phase shifts from coral to macroalgal dominance on coral reefs. *Ecology* **90**, 1478–1484 (2009). doi: 10.1890/08-1781.1; pmid: 19569362
48. T. A. Gardner, I. M. Côté, J. A. Gill, A. Grant, A. R. Watkinson, Long-term region-wide declines in Caribbean corals. *Science* **301**, 958–960 (2003). doi: 10.1126/science.1086050; pmid: 12869698
49. J. S. Harding, E. F. Benfield, P. V. Bolstad, G. S. Helfman, E. B. Jones 3rd, Stream biodiversity: The ghost of land use past. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14843–14847 (1998). doi: 10.1073/pnas.95.25.14843; pmid: 9843977
50. B. Sinervo, C. M. Lively, The rock-paper-scissors game and the evolution of alternative male strategies. *Nature* **380**, 240–243 (1996). doi: 10.1038/380240a0
51. L. Buss, J. Jackson, Competitive networks: Nontransitive competitive relationships in cryptic coral reef environments. *Am. Nat.* **113**, 223–234 (1979). doi: 10.1086/283381

52. R. Marrs, M. J. Hicks, Study of vegetation change at lakenheath warren: A re-examination of A. S. Watt's theories of bracken dynamics in relation to succession and vegetation management. *J. Appl. Ecol.* **23**, 1029 (1986). doi: 10.2307/2403953
53. K. Higgins, A. Hastings, J. Sarvela, L. Botsford, Stochastic dynamics and deterministic skeletons: Population behavior of Dungeness crab. *Science* **276**, 1431–1435 (1997). doi: 10.1126/science.276.5317.1431
54. B. Dennis, R. A. Desharnais, J. Cushing, S. M. Henson, R. Costantino, Can noise induce chaos? *Oikos* **102**, 329–339 (2003). doi: 10.1034/j.1600-0706.2003.12387.x
55. J. Huisman, F. J. Weissing, Biodiversity of plankton by species oscillations and chaos. *Nature* **402**, 407–410 (1999). doi: 10.1038/46540
56. T. Fukami, Assembly history interacts with ecosystem size to influence species diversity. *Ecology* **85**, 3234–3242 (2004). doi: 10.1890/04-0340
57. T. Fukami, T. M. Bezemer, S. R. Mortimer, W. H. van der Putten, Species divergence and trait convergence in experimental plant community assembly. *Ecol. Lett.* **8**, 1283–1290 (2005). doi: 10.1111/j.1461-0248.2005.00829.x
58. C. R. Bulman *et al.*, Minimum viable metapopulation size, extinction debt, and the conservation of a declining species. *Ecol. Appl.* **17**, 1460–1473 (2007). doi: 10.1890/06-1032.1; pmid: 17708221
59. R. A. Ims, J. A. Henden, S. T. Killengreen, Collapsing population cycles. *Trends Ecol. Evol.* **23**, 79–86 (2008). doi: 10.1016/j.tree.2007.10.010; pmid: 18191281
60. M. Kuussaari *et al.*, Extinction debt: A challenge for biodiversity conservation. *Trends Ecol. Evol.* **24**, 564–571 (2009). doi: 10.1016/j.tree.2009.04.011; pmid: 19665254
61. J. MacHunter, W. Wright, R. Loyn, P. Rayment, Bird declines over 22 years in forest remnants in southeastern Australia: Evidence of faunal relaxation? *Can. J. For. Res.* **36**, 2756–2768 (2006). doi: 10.1139/x06-159
62. E. Candès, J. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.* **59**, 1207–1223 (2006). doi: 10.1002/cpa.20124
63. E. Candès, in *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), vol. 3, pp. 1433–1452.
64. R. G. Baraniuk, Compressed sensing. *IEEE Signal Process. Mag.* **24**, 118–121 (2007). doi: 10.1109/MSP.2007.4286571
65. E. Candès, M. Wakin, An introduction to compressive sampling. *IEEE Signal Process. Mag.* **25**, 21–30 (2008). doi: 10.1109/MSP.2007.914731
66. T. P. Hughes, C. Linares, V. Dakos, I. A. van de Leemput, E. H. van Nes, Living dangerously on borrowed time during slow, unrecognized regime shifts. *Trends Ecol. Evol.* **28**, 149–155 (2013). doi: 10.1016/j.tree.2012.08.022; pmid: 22995893
67. B. Hörnfeldt, Long-term decline in numbers of cyclic voles in boreal Sweden: Analysis and presentation of hypotheses. *Oikos* **107**, 376–392 (2004). doi: 10.1111/j.0030-1299.2004.13348.x
68. NERC Centre for Population Biology, Imperial College, Global Population Dynamics Database (1999); www3.imperial.ac.uk/cpb/research/patternsandprocesses/gpdd.
69. C. Moore, C. Stieha, B. Nolting, M. K. Cameron, K. C. Abbott, QPot: An R package for calculating quasi-potentials (2015); https://cran.r-project.org/web/packages/QPot/index.html.
70. S. Strogatz, *Nonlinear Dynamics and Chaos with Applications to Physics, Biology, Chemistry, and Engineering* (Westview Press, 1994).
71. M. Kot, *Elements of Mathmatical Ecology* (Cambridge Univ. Press, 2001).
72. U. Sommer, Comparison between steady-state and nonsteady state competition: Experiments with natural phytoplankton. *Limnol. Oceanogr.* **30**, 335–346 (1985). doi: 10.4319/lo.1985.30.2.0335

ACKNOWLEDGMENTS

We thank S. Catella, R. Decker, L. Gross, F. Ji, B. Lerch, A. Patterson, C. Plunkett, R. Snyder, A. Strang, and E. White for helpful feedback on an earlier draft. **Funding:** This work resulted from a working group supported by the National Institute for Mathematical Biological Synthesis, a synthesis center sponsored by the National Science Foundation through award DBI-1300426 with additional support from the University of Tennessee, Knoxville. **Competing interests:** There are no competing interests.

10.1126/science.aat6412

RESEARCH ARTICLE SUMMARY

NEURODEGENERATION

Combined adult neurogenesis and BDNF mimic exercise effects on cognition in an Alzheimer's mouse model

Se Hoon Choi, Enjana Bylykbashi, Zena K. Chatila, Star W. Lee, Benjamin Pulli, Gregory D. Clemenson, Eunhee Kim, Alexander Rompala, Mary K. Oram, Caroline Asselin, Jenna Aronson, Can Zhang, Sean J. Miller, Andrea Lesinski, John W. Chen, Doo Yeon Kim, Henriette van Praag, Bruce M. Spiegelman, Fred H. Gage, Rudolph E. Tanzi*

INTRODUCTION: Alzheimer's disease (AD) is the most common form of age-related dementia, characterized by cognitive impairment, neurodegeneration, β-amyloid (Aβ) deposition, neurofibrillary tangle formation, and neuroinflammation. The most popular therapeutic approach aimed at reducing Aβ burden has not yet proved effective in halting disease progression. A successful therapy would both remove the pathological hallmarks of the disease and provide some functional recovery. The hippocampus contains neural progenitor cells that continue to generate new neurons, a process called adult hippocampal neurogenesis (AHN). AHN is impaired before the onset of classical AD pathology in AD mouse models. Human AHN has also been reported to be altered in AD patients. However, evidence supporting a role for AHN in AD has remained sparse and inconclusive.

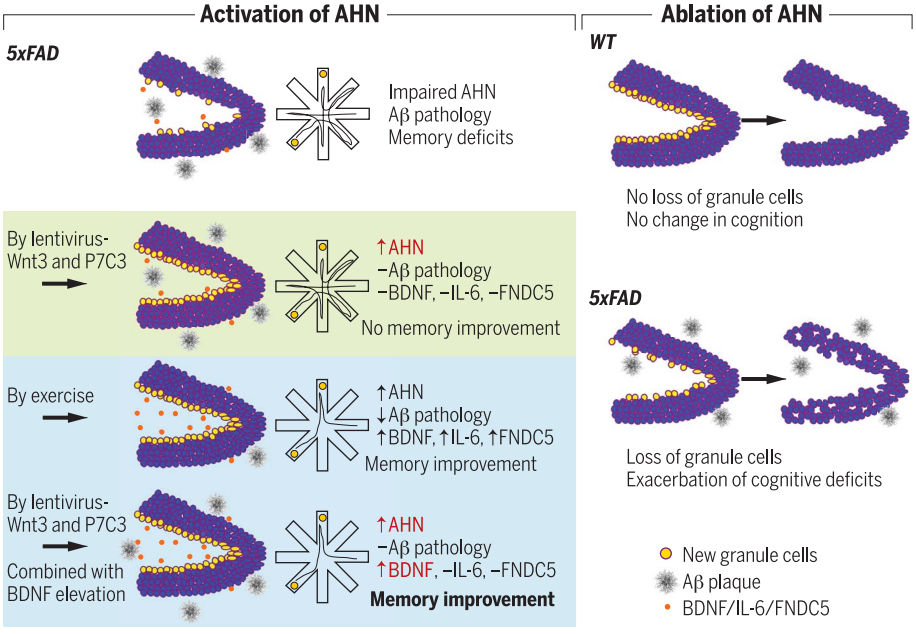
RATIONALE: Two fundamental questions remain: (i) whether AHN could be enhanced and exploited for therapeutic purposes for AD, and (ii) whether AHN impairment mediates aspects of AD pathogenesis. To address these questions, we increased AHN genetically (WNT3) and pharmacologically (P7C3) in AD transgenic 5×FAD mice and explored whether promoting AHN alone can ameliorate AD pathology and behavioral symptoms. We assessed the role of exercise, a known neurogenic stimulus, and explored whether promoting AHN in conjunction with the salutary biochemical changes induced by exercise can improve AD pathology and behavioral symptoms in mice. We also investigated whether AHN suppression, by irradiation, temozolomide, or dominant-negative WNT, contributes to AD pathogenesis and assessed the functional roles of AHN in AD.

RESULTS: Inducing AHN alone conferred minimal to no benefit for improving cognition in 5×FAD mice. Exercise-induced AHN improved cognition along with reduced Aβ load and increased levels of brain-derived neurotrophic factor (BDNF), interleukin-6 (IL-6), fibronectin type III domain–containing protein–5 (FNDC5), and synaptic markers. However, AHN activation was also required for exercise-induced improvement in memory. Inducing AHN genetically and pharmacologically in combination with elevating BDNF levels mimicked beneficial effects of exercise on AD mice. Conversely, suppressing AHN in early stages of AD exacerbated neuronal vulnerability in later stages of AD, leading to cognitive impairment and increased neuronal loss. However, no such effects from AHN ablation were observed in nontransgenic wild-type (WT) mice, suggesting that AHN has a specific role in AD.

CONCLUSION: Promoting AHN can only ameliorate AD pathology and cognitive deficits in the presence of a healthier, improved local brain environment, e.g., stimulated by exercise. Increasing AHN alone combined with overexpression of BDNF could mimic exercise-induced improvements in cognition, without reducing Aβ burden. Adult-born neurons generated very early in life are critical for maintaining hippocampal neuronal populations in the hostile brain environment created by AD later in life. Thus, AHN impairment may be a primary event that later mediates other aspects of AD pathogenesis. Future attempts to create pharmacological mimetics of the benefits of exercise on both increased AHN and BDNF may someday provide an effective means for improving cognition in AD. Moreover, increasing neurogenesis in the earliest stages of AD pathogenesis may protect against neuronal cell death later in the disease, providing a potentially powerful disease-modifying treatment strategy for AD. ■

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aan8821



Role of adult-born neurons in AD. Inducing AHN alone by WNT3 and P7C3 together did not prevent cognitive dysfunction, whereas activating AHN through exercise improved memory in 5×FAD mice. Increasing AHN alone together with overexpression of BDNF could mimic exercise-induced improvement in cognition. Suppressing AHN exacerbated neuronal vulnerability, leading to cognitive impairment and increased neuronal loss in 5×FAD mice, but not in WT mice.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: tanzi@helix.mgh.harvard.edu
Cite this article as S. H. Choi *et al.*, *Science* **361**, eaan8821 (2018). DOI: 10.1126/science.aan8821

RESEARCH ARTICLE

NEURODEGENERATION

Combined adult neurogenesis and BDNF mimic exercise effects on cognition in an Alzheimer's mouse model

Se Hoon Choi[1], Enjana Bylykbashi[1], Zena K. Chatila[1]*, Star W. Lee[2]†, Benjamin Pulli[3], Gregory D. Clemenson[2]‡, Eunhee Kim[1], Alexander Rompala[1], Mary K. Oram[1], Caroline Asselin[1], Jenna Aronson[1]§, Can Zhang[1], Sean J. Miller[1]||, Andrea Lesinski[1]¶, John W. Chen[3], Doo Yeon Kim[1], Henriette van Praag[4], Bruce M. Spiegelman[5], Fred H. Gage[2], Rudolph E. Tanzi[1]#

Adult hippocampal neurogenesis (AHN) is impaired before the onset of Alzheimer's disease (AD) pathology. We found that exercise provided cognitive benefit to 5×FAD mice, a mouse model of AD, by inducing AHN and elevating levels of brain-derived neurotrophic factor (BDNF). Neither stimulation of AHN alone, nor exercise, in the absence of increased AHN, ameliorated cognition. We successfully mimicked the beneficial effects of exercise on AD mice by genetically and pharmacologically inducing AHN in combination with elevating BDNF levels. Suppressing AHN later led to worsened cognitive performance and loss of preexisting dentate neurons. Thus, pharmacological mimetics of exercise, enhancing AHN and elevating BDNF levels, may improve cognition in AD. Furthermore, applied at early stages of AD, these mimetics may protect against subsequent neuronal cell death.

Alzheimer's disease (AD) is the most common form of age-related dementia, characterized by cognitive impairment, neurodegeneration, deposition of β-amyloid (Aβ), neurofibrillary tangle formation, and neuroinflammation (*1*). The most popular therapeutic approach aimed at reducing Aβ burden has not yet proved effective in halting disease progression. A successful therapy would ideally both remove the pathological hallmarks of the disease and provide a level of functional recovery.

[1]Genetics and Aging Research Unit, Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA 02129, USA. [2]Laboratoy of Genetics, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA. [3]Institute for Innovation in Imaging, Department of Radiology, Massachusetts General Hospital, Boston, MA 02114, USA. [4]Department of Biomedical Science, Charles E. Schmidt College of Medicine, and Brain Institute, Florida Atlantic University, Jupiter, FL 33458, USA. [5]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02115, USA.
*Present address: Department of Neuroscience, Columbia University, New York, NY 10036, USA. †Present address: Department of Evolution, Ecology, and Organismal Biology, University of California, Riverside, Riverside, CA 92521, USA. ‡Present address: Department of Neurobiology and Behavior, University of California, Irvine, Irvine, CA 92697, USA. §Present address: Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. ||Present address: Department of Neurology and Neurological Sciences, School of Medicine, Stanford University, Stanford, CA 94305, USA. ¶Present address: Cell and Molecular Biology, University of Rhode Island, Kingston, RI 02881, USA.
#Corresponding author. Email: tanzi@helix.mgh.harvard.edu

The human hippocampus contains neural progenitor cells (NPCs) that continue to generate new neurons, a process called adult hippocampal neurogenesis (AHN) (*2*). Although adult-generated neurons play an important role in learning and memory under physiological conditions, their function under pathological conditions, such as those of AD, has remained elusive. Emerging evidence indicates that AHN is impaired prior to the onset of classical AD pathology in AD mouse models, e.g., Aβ deposition (*3*). Human AHN has also been reported to be altered in AD patients (*4*–*9*).

To date, however, the evidence supporting a role for AHN in AD has remained sparse and inconclusive. Two fundamental questions remain: (i) whether AHN could be enhanced and exploited for therapeutic purposes for AD, and (ii) whether AHN impairment mediates aspects of AD pathogenesis or is a neuroadaptive response to the complex pathological events of the disease. We set out to address these two questions genetically and pharmacologically in 5×FAD mice (*10*). We also assessed the role of physical exercise, a known neurogenic stimulus that counteracts various aspects of AD pathology (*11*–*13*), and explored whether promoting AHN in conjunction with the salutary biochemical changes that are induced by exercise can ameliorate AD pathology and behavioral symptoms in mice. Finally, we investigated how impairment of AHN contributes to AD pathogenesis and assessed the functional roles of adult-generated neurons in the pathological course of AD.

AHN stimulation with LV-Wnt3 and P7C3 or with exercise

To stimulate AHN, beginning at 2 months of age for 4 or 4.5 months, sedentary 5×FAD mice were injected with P7C3, a compound that enhances NPC survival (*14*). At the 3-month time point, these mice also received lentivirus expressing WNT3 protein (LV-Wnt3) to increase NPC proliferation (*15*) (Fig. 1, A and B). Control 5×FAD mice were injected with vehicle and lentivirus expressing green fluorescent protein (LV-GFP, referred to as "5×FAD^{CTL}" mice). As an alternative means of inducing AHN, we also tested the effects of exercise on a 5×FAD mice cohort. Successful promotion of AHN with P7C3 and LV-Wnt3 or with exercise was observed in the male and female 5×FAD mice, as determined by immunostaining for doublecortin (DCX)$^+$ neurons (Fig. 1, C and D). After completing cognition tasks, mice with similar DCX$^+$ cell counts among the 5×FAD mice treated with P7C3 and LV-Wnt3 and the exercised 5×FAD mice were selected for further study; we included all those with levels higher than the maximum level seen in 5×FAD^{CTL} mice (mice in dashed box in Fig. 1D; see table S1 for animal numbers in each experimental group and group arrangement explanations). From here on, we refer to the subsets of 5×FAD mice treated with P7C3 and LV-Wnt3 to promote AHN or exercised 5×FAD mice in the box in Fig. 1D as "5×FAD^{ProAHN}" and "5×FAD$^{+AHN(RUN)}$" mice, respectively. Exercised 5×FAD mice with less than the maximum DCX$^+$ cell count seen in 5×FAD^{CTL} mice were designated "5×FAD$^{\psi AHN(RUN)}$" mice. Mice were sacrificed at the age of 6 to 6.5 months, when untreated 5×FAD mice showed cognitive deficits compared to nontransgenic wild-type mice (WT) mice (fig. S1). Endogenous neurogenic changes and AD pathologies observed in untreated 5×FAD mice, including our rationale for employing this mouse line, are described in figs. S2 and S3 and materials and methods.

We tested whether AHN stimulation by exercise versus P7C3 and LV-Wnt3 affected Aβ plaque amounts. Immunostaining with anti-Aβ antibody 3D6 showed that, whereas exercised mice exhibited a decreased Aβ burden, activation of AHN alone (by P7C3 and LV-Wnt3) did not change Aβ plaque amounts (Fig. 1, C and E). 5×FAD$^{\psi AHN(RUN)}$ mice also had a reduced Aβ burden, suggesting that an effect due to exercise and not AHN alone contributed to changes in Aβ burden after physical activity.

Increasing AHN alone did not ameliorate cognitive function in 5×FAD mice

We examined whether promoting AHN could ameliorate cognitive impairment in AD. We performed a delayed nonmatching to place (DNMP) task to measure spatial pattern separation, an eight-arm radial arm maze (RAM) to measure reference and retention memory, and a Y-maze to measure spatial working memory. Pattern separation, which is the formation of distinct

representations of similar inputs, has been shown to require adult-generated neurons (*16*). Reduced ability to separate patterns (i.e., to recognize differences between very similar events) is one of the first behavioral deficits in patients with mild cognitive impairments, which often progress to AD (*17*). The other cognition types listed are well accepted to be hippocampus dependent; however, it remains unclear whether AHN contributes to these memory types. We used the RAM and Y-maze, which are routine methods to assess memory function in AD mice (*18*), to explore whether increasing AHN also ameliorates AD-associated cognitive impairment in 5×FAD mice.

In the DNMP task, male cohorts of 5×FAD^{ProAHN} and 5×FAD$^{+AHN(RUN)}$ mice showed similar improvements in pattern separation compared to 5×FAD^{CTL} mice at 5.5 to 6 months of age (Fig. 2A, left graph). However, in female cohorts, 5×FAD^{ProAHN} mice failed to show improved pattern separation memory compared to 5×FAD^{CTL} mice, whereas 5×FAD$^{+AHN(RUN)}$ mice showed improvement (Fig. 2A, right graph). Both male and female 5×FAD^{ProAHN} mice failed to show improved working memory in the Y-maze, whereas 5×FAD$^{+AHN(RUN)}$ mice performed better (Fig. 2B and fig. S4). Male mice were tested in the RAM task. In the training trials of the task, all the groups showed a clear learning curve during the 5-day training session (Fig. 2C, left graph). However, whereas 5×FAD$^{+AHN(RUN)}$ mice showed significantly improved reference memory compared to 5×FAD^{CTL} mice on days 2 and 3, 5×FAD^{ProAHN} mice failed to show improved memory. The performance of 5×FAD^{ProAHN} mice did not differ from that of 5×FAD^{CTL} mice. In the retention memory task, on day 8, 5×FAD^{ProAHN} mice again failed to show improved memory, whereas 5×FAD$^{+AHN(RUN)}$ mice did (Fig. 2C, right graph). Our results suggest that increasing AHN alone was sufficient to improve pattern separation memory in male 5×FAD mice but not in female mice; however, it was not sufficient to improve other forms of cognition. By contrast, exercise

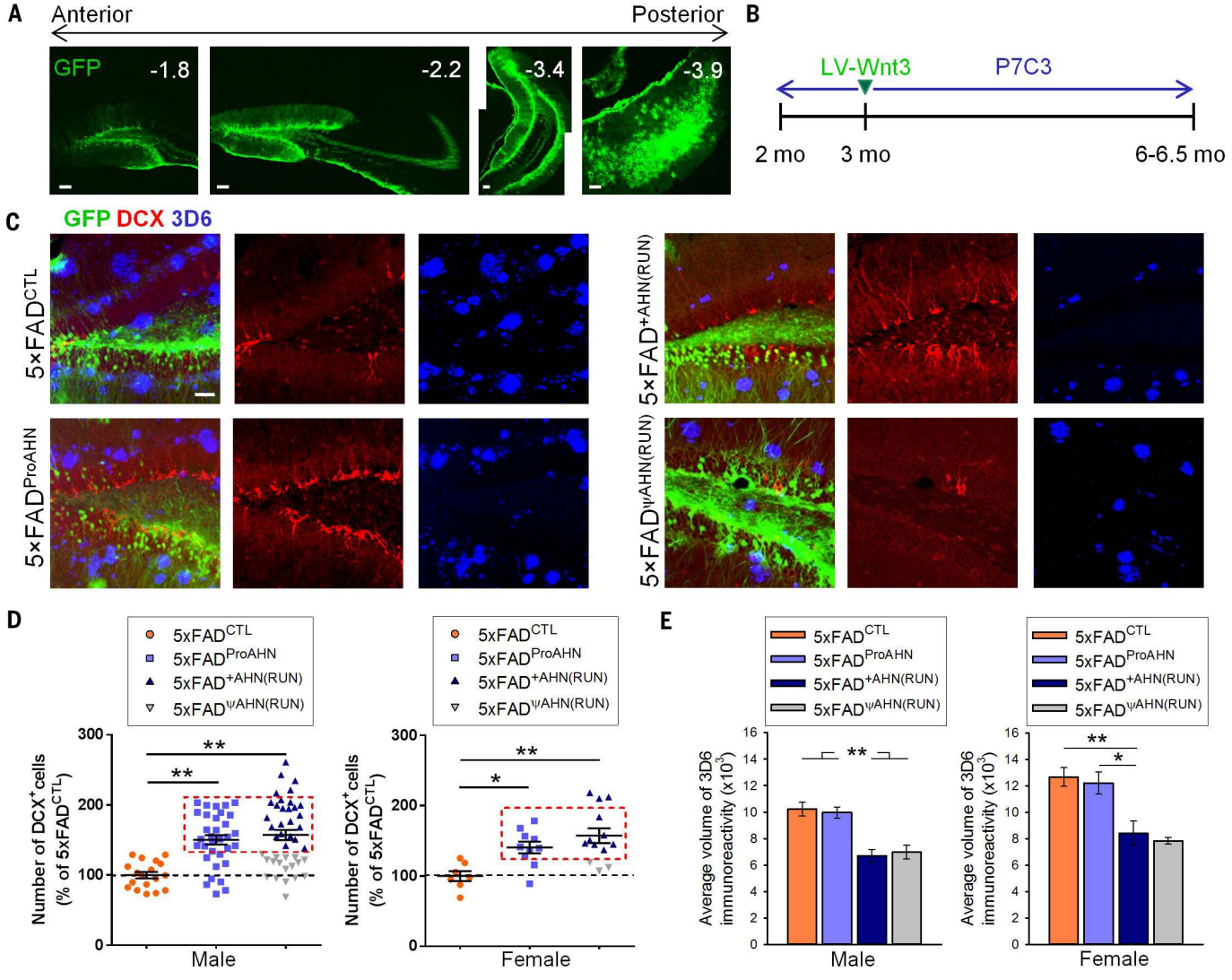


Fig. 1. AHN activation alone does not change Aβ plaque levels. (**A**) Stereotaxic injection of lentiviral vectors targets adult DG. Numbers (upper right) are approximate distances from bregma. Scale bar: 100 μm. (**B**) Experimental procedures timeline. (**C**) Photomicrographs of DCX$^+$ cells and 3D6$^+$ Aβ plaques in the hippocampus of 5×FAD^{CTL}, 5×FAD^{ProAHN}, 5×FAD$^{+AHN(RUN)}$, and 5×FAD$^{ψAHN(RUN)}$ mice. Scale bar: 50 μm. (**D**) Distribution of AHN activation by P7C3 along with LV-Wnt3 or by exercise. Data points represent DCX$^+$ cell count per mouse, as percentage of mean DCX$^+$ cell count for 5×FAD^{CTL} mice by gender. In male 5×FAD^{CTL}, 5×FAD^{ProAHN}, and 5×FAD^{RUN} mice, $F_{(2,90)} = 13.57$, $P < 0.01$. In female, $F_{(2,27)} = 8.05$, $P < 0.01$. (**E**) Quantitative analysis of Aβ burden volume in the hippocampus of 5×FAD^{CTL}, 5×FAD^{ProAHN}, 5×FAD$^{+AHN(RUN)}$, and 5×FAD$^{ψAHN(RUN)}$ mice. Volume in arbitrary units (mean voxel count ± SEM; male, $F_{(3,76)} = 15.57$, $P < 0.01$; female, $F_{(2,19)} = 7.659$, $P < 0.01$). Female 5×FAD$^{ψAHN(RUN)}$ mice were excluded due to low *n*.

led to improved cognitive function in all three types of memory tests.

Interestingly, exercise alone, in the absence of increased AHN, exerted no observed beneficial effects on cognitive function ($5\times FAD^{\psi AHN(RUN)}$, Fig. 2). This finding suggests that increased AHN above physiological levels is required for the positive behavioral effects of exercise in 5×FAD mice. However, a lack of AHN in WT mice does not block the beneficial behavioral effects of increased physical activity gained through living in an enriched environment (*19*). We hypothesized that adult-generated neurons have specific functions in AD pathological conditions and in mediating the effects of exercise in AD. To test these points, we assessed the effects of AHN ablation in 5×FAD mice versus WT mice.

Ablating AHN induced cell death in dentate gyrus (DG) of 5×FAD mice

To block AHN, 6- to 8-week-old male WT and 5×FAD mice received either focal irradiation (*20*) (IR; WT^{IR}; $5\times FAD^{IR}$), the DNA-alkylating agent temozolomide (*21*) (TMZ; WT^{TMZ}; $5\times FAD^{TMZ}$), or a lentivirus expressing a dominant-negative form of WNT (*15*) (LV-dnWnt; $WT^{LV\text{-}dnWnt}$; $5\times FAD^{LV\text{-}dnWnt}$) (Fig. 3, A to C). They were then sacrificed at 3 or 5 months of age. Sham mice (WT^{Sham}; $5\times FAD^{Sham}$) and mice treated with vehicle (WT^{Veh}; $5\times FAD^{Veh}$) or LV-GFP ($WT^{LV\text{-}GFP}$; $5\times FAD^{LV\text{-}GFP}$) served as controls for mice treated with IR, TMZ, and LV-dnWnt, respectively. IR almost completely eliminated AHN, as determined by immunostaining for DCX^+ neurons, in both WT and 5×FAD mice. Injection of TMZ or LV-dnWnt also showed significant, although variable, AHN knockdown (Fig. 3C; see table S2 for animal numbers in each experimental group and group arrangement explanations).

To test whether adult-generated neurons are required to maintain stability in existing neuronal populations in AD, we stained brain sections from each group with an antibody against activated Caspase 3 (Casp3), a marker for apoptotic cells, and examined them for evidence of cell loss. $5\times FAD^{Sham}$, $5\times FAD^{Veh}$, and $5\times FAD^{LV\text{-}GFP}$ mice showed only a few $Casp3^+$ cells in the DG at 5 months of age. Conversely, the number of $Casp3^+$ cells was significantly increased in the DG of 5-month-old $5\times FAD^{IR}$, $5\times FAD^{TMZ}$, and $5\times FAD^{LV\text{-}dnWnt}$ mice; no such cell death was observed in the corresponding WT treatment groups (Fig. 3, D and E). However, $5\times FAD^{TMZ}$ and $5\times FAD^{LV\text{-}dnWnt}$ mice that showed less than ~60% AHN reduction did not exhibit a significant increase in the number of $Casp3^+$ cells compared to $5\times FAD^{Veh}$ and $5\times FAD^{LV\text{-}GFP}$ mice, respectively (Fig. 3F). These results suggest that the existence of $Casp3^+$ cells in $5\times FAD^{TMZ}$ and $5\times FAD^{LV\text{-}dnWnt}$ mice could be dependent on the degree of AHN knockdown.

Therefore, we regrouped $5\times FAD^{TMZ}$ mice based on the degree of AHN knockdown. $5\times FAD^{TMZ\ (Mod\ KD)}$ showed moderate (less than 60% AHN reduction) and $5\times FAD^{TMZ\ (high\ KD)}$ showed high (over 60% AHN reduction) AHN knockdown. Likewise, the $5\times FAD^{LV\text{-}dnWnt}$ group was regrouped into $5\times FAD^{LV\text{-}dnWnt\ (Mod\ KD)}$ and $5\times FAD^{LV\text{-}dnWnt\ (high\ KD)}$ mice. WT mice injected with TMZ or LV-dnWnt

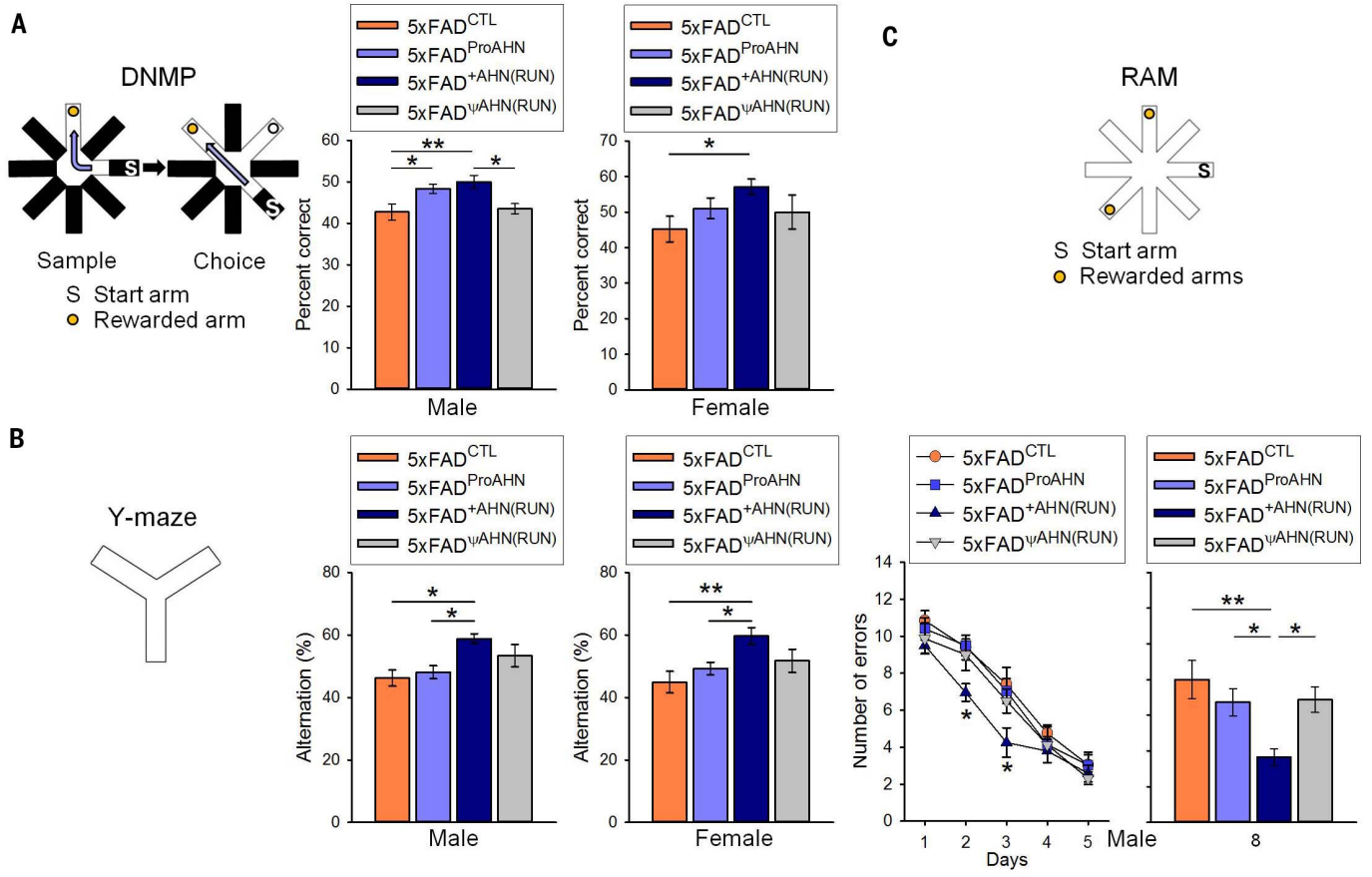


Fig. 2. Increasing AHN alone does not ameliorate cognitive function in 5×FAD mice, whereas exercise-induced AHN does. (**A**) Schematic of DNMP in the RAM task, which consisted of sample and choice phase, and quantification of percent correct during choice phase (male, $F_{(3,31)}$ = 5.983, $P < 0.01$; female, $F_{(2,19)}$ = 3.887, $P < 0.05$). (**B**) Schematic of Y-maze task and spontaneous alternation behavior (male, $F_{(3,31)}$ = 3.935, $P < 0.05$; female, $F_{(2,19)}$ = 7.416, $P < 0.01$). Total arm entries were comparable among groups (fig. S4). (**C**) Schematic of RAM task and mean error number in training trials (left graph). Two-way ANOVA with repeated measures revealed significant effects for days ($F_{(4,164)}$ = 99.29, $P < 0.01$) and groups ($F_{(3,41)}$ = 5.129, $P < 0.01$) but not interaction ($F_{(12,164)}$ = 0.9894, P = 0.4613). Analysis of error number on each day by Fisher's LSD post hoc tests revealed $5\times FAD^{+AHN(RUN)}$ differed significantly from both $5\times FAD^{CTL}$ and $5\times FAD^{ProAHN}$ mice on days 2 and 3 (day 2, $F_{(3,41)}$ = 4.074, $P < 0.05$; day 3, $F_{(3,41)}$ = 3.499, $P < 0.05$). Right graph: mean error number in memory retention trial ($F_{(3,41)}$ = 5.675, $P < 0.01$).

were also regrouped into $WT^{TMZ\ (Mod\ KD)}$ and $WT^{TMZ\ (high\ KD)}$ groups, and $WT^{LV\text{-}dnWnt\ (Mod\ KD)}$ and $WT^{LV\text{-}dnWnt\ (high\ KD)}$ groups. As shown in Fig. 3F, $5\times FAD^{TMZ\ (high\ KD)}$ mice showed significantly increased $Casp3^+$ cell numbers compared to both $5\times FAD^{Veh}$ and $5\times FAD^{TMZ\ (Mod\ KD)}$ mice, whereas $5\times FAD^{TMZ\ (Mod\ KD)}$ mice did not show increased $Casp3^+$ cells compared to $5\times FAD^{Veh}$ mice. Similarly, $5\times FAD^{LV\text{-}dnWnt\ (high\ KD)}$ mice showed increased $Casp3^+$ cell numbers compared to both $5\times FAD^{LV\text{-}GFP}$ and $5\times FAD^{LV\text{-}dnWnt\ (Mod\ KD)}$ mice, whereas no significant difference was observed between $5\times FAD^{LV\text{-}GFP}$ and $5\times FAD^{LV\text{-}dnWnt\ (Mod\ KD)}$ mice.

From here on, $5\times FAD^{Sham}$, $5\times FAD^{Veh}$, and $5\times FAD^{LV\text{-}GFP}$ mice are collectively referred to as $5\times FAD^{CTL}$; $5\times FAD^{IR}$, $5\times FAD^{TMZ}$, and $5\times FAD^{LV\text{-}dnWnt}$ mice as $5\times FAD^{\text{-}AHN}$ mice; $5\times FAD^{IR}$, $5\times FAD^{TMZ\ (High\ KD)}$, and $5\times FAD^{LV\text{-}dnWnt\ (High\ KD)}$ mice as $5\times FAD^{\text{-}AHN\ (High\ KD)}$ mice; $5\times FAD^{TMZ\ (Mod\ KD)}$ and $5\times FAD^{LV\text{-}dnWnt\ (Mod\ KD)}$ mice as $5\times FAD^{\text{-}AHN\ (Mod\ KD)}$ mice. Likewise, the corresponding groups of WT mice are referred to as WT^{CTL}, $WT^{\text{-}AHN}$, $WT^{\text{-}AHN\ (High\ KD)}$, and $WT^{\text{-}AHN\ (Mod\ KD)}$ mice, respectively.

Most $Casp3^+$ cells were colabeled with NeuN, a mature neuronal marker (Fig. 4A), suggesting that AHN suppression caused the death of mature neurons. The total number of granule cells was decreased by the death of mature neurons in the $5\times FAD^{\text{-}AHN\ (High\ KD)}$ group but not in the $WT^{\text{-}AHN\ (High\ KD)}$ group (Fig. 4B), demonstrating the importance of AHN in the survival of the preexisting granule cell population. The number of granule cells was not decreased in $5\times FAD^{\text{-}AHN\ (Mod\ KD)}$ mice. Our results suggest that AHN plays a potential function in maintaining

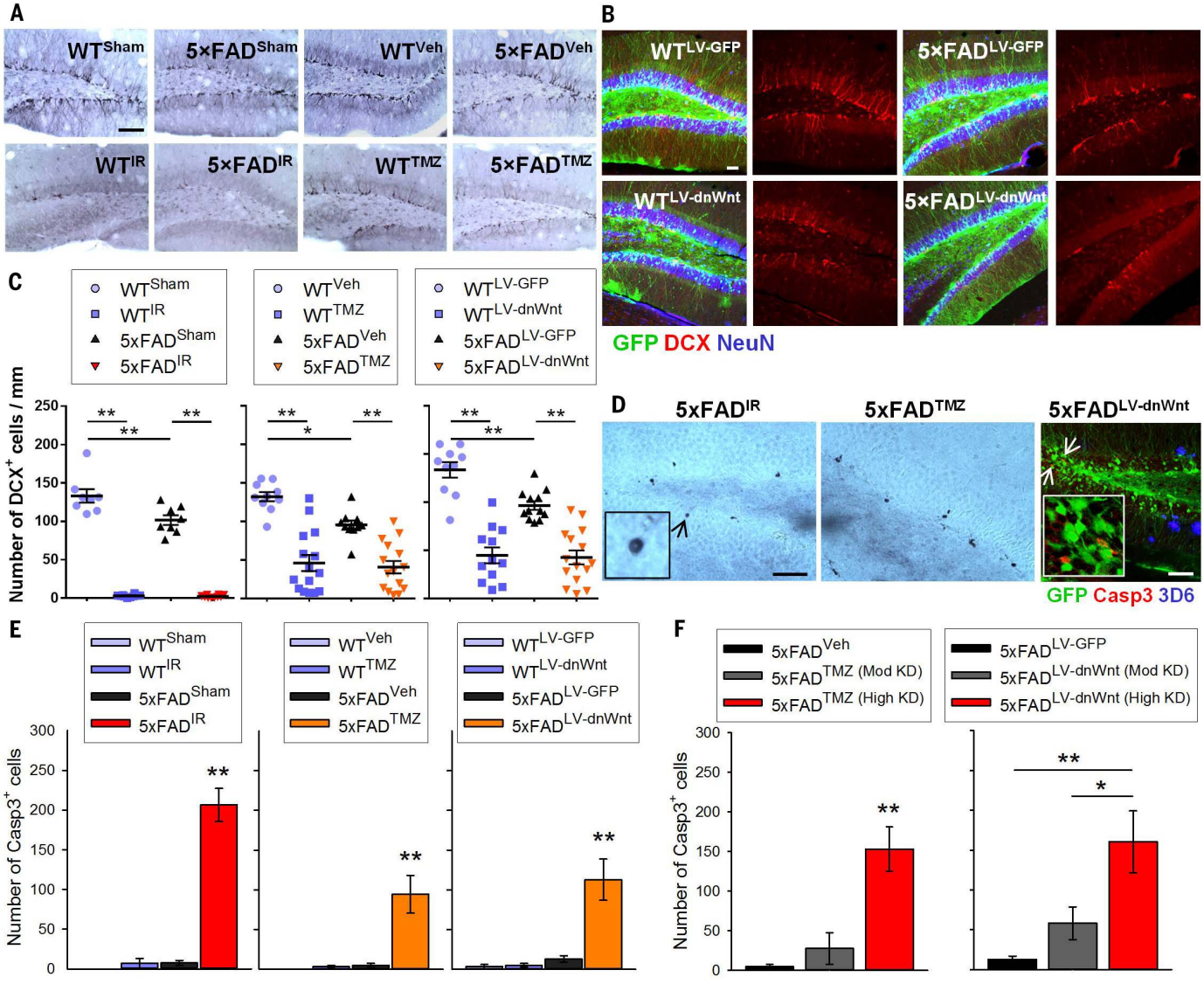


Fig. 3. Ablating AHN induces cell death in 5×FAD mice. (**A**) Photomicrographs of DCX^+ cells in the DG of 5-month-old WT^{Sham}, $5\times FAD^{Sham}$, WT^{Veh}, $5\times FAD^{Veh}$, WT^{IR}, $5\times FAD^{IR}$, WT^{TMZ}, and $5\times FAD^{TMZ}$ mice. Scale bar: 100 μm. (**B**) Photomicrographs of DCX^+ cells in the transduced DG of WT and 5×FAD mice by LV-GFP or LV-dnWnt. Mature granule neurons are stained for NeuN. Scale bar: 50 μm. (**C**) Quantification of DCX^+ cells in male WT^{Sham}, WT^{IR}, $5\times FAD^{Sham}$, and $5\times FAD^{IR}$ ($F_{(3,32)} = 197.9$, $P < 0.01$; left), WT^{Veh}, WT^{TMZ}, $5\times FAD^{Veh}$, and $5\times FAD^{TMZ}$ ($F_{(3,46)} = 23.96$, $P < 0.01$; middle), and $WT^{LV\text{-}GFP}$, $WT^{LV\text{-}dnWnt}$, $5\times FAD^{LV\text{-}GFP}$, and $5\times FAD^{LV\text{-}dnWnt}$ mice ($F_{(3,45)} = 35.21$, $P < 0.01$; right). (**D**) Representative images of $Casp3^+$ cells in $5\times FAD^{IR}$ (left), $5\times FAD^{TMZ}$ (middle), or $5\times FAD^{LV\text{-}dnWnt}$ (right) mice. Insets represent digital magnification of arrow-indicated $Casp3^+$ cells. Scale bars: 50 μm. (**E**) Quantification of $Casp3^+$ cells in WT^{Sham}, WT^{IR}, $5\times FAD^{Sham}$, and $5\times FAD^{IR}$ ($F_{(3,32)} = 72.38$, $P < 0.01$; left), WT^{Veh}, WT^{TMZ}, $5\times FAD^{Veh}$, and $5\times FAD^{TMZ}$ ($F_{(3,46)} = 11.26$, $P < 0.01$; middle), and $WT^{LV\text{-}GFP}$, $WT^{LV\text{-}dnWnt}$, $5\times FAD^{LV\text{-}GFP}$, and $5\times FAD^{LV\text{-}dnWnt}$ mice ($F_{(3,45)} = 11.98$, $P < 0.01$; right). (**F**) Quantification of $Casp3^+$ cells in $5\times FAD^{Veh}$, $5\times FAD^{TMZ\ (Mod\ KD)}$, and $5\times FAD^{TMZ\ (High\ KD)}$ ($F_{(2,22)} = 19.41$, $P < 0.01$; left), and $5\times FAD^{LV\text{-}GFP}$, $5\times FAD^{LV\text{-}dnWnt\ (Mod\ KD)}$, and $5\times FAD^{LV\text{-}dnWnt\ (High\ KD)}$ mice ($F_{(2,24)} = 12.17$, $P < 0.01$; right).

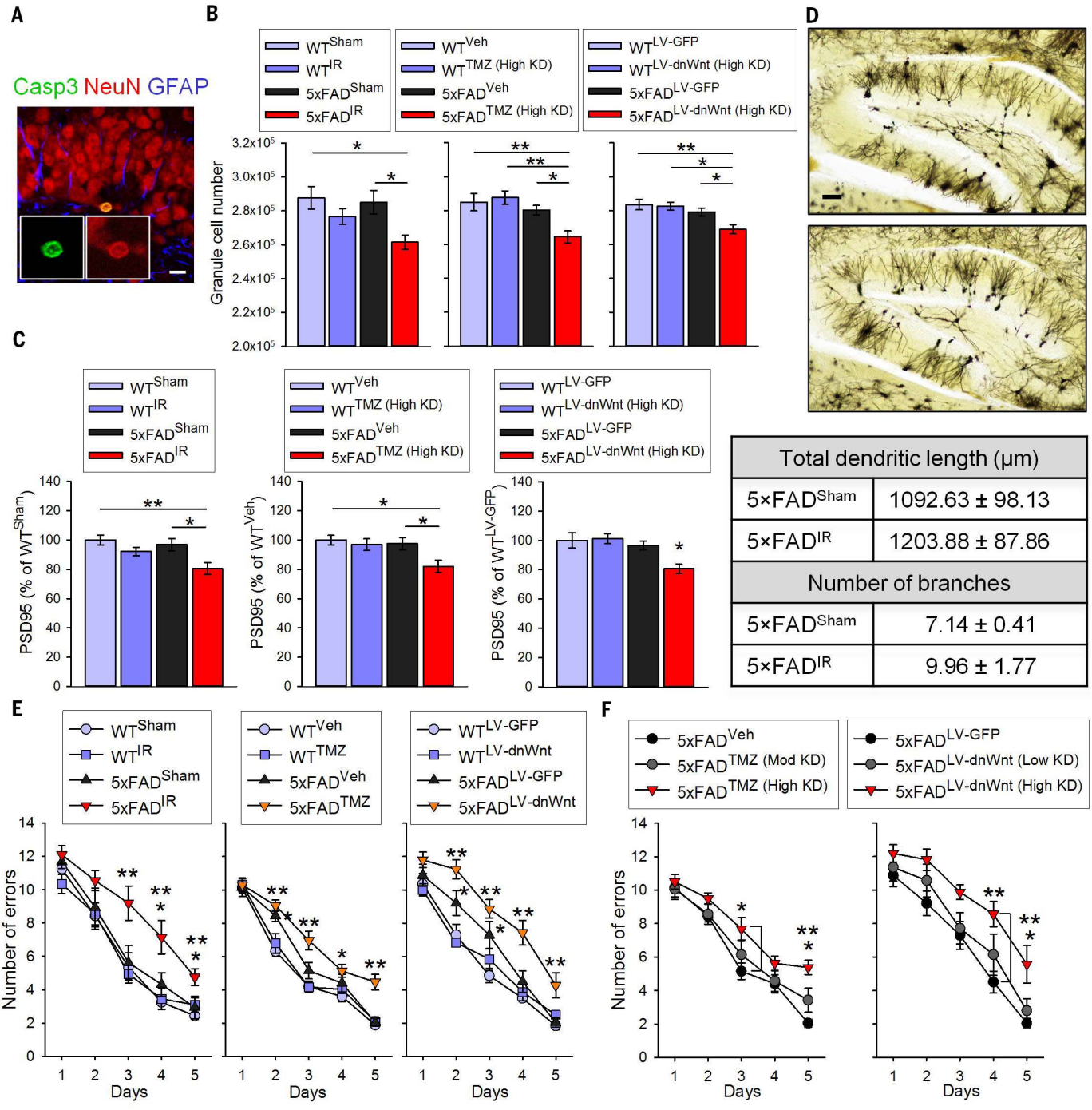


Fig. 4. Ablating AHN induces granule cell and synaptic marker loss, and exacerbates cognitive impairment in 5×FAD mice. (**A**) Representative images of Casp3$^+$ cell colabeled with NeuN. Scale bar: 10 μm. (**B**) Quantification of granule cell number in WT^{Sham}, WT^{IR}, 5×FAD^{Sham}, and 5×FAD^{IR} mice ($F_{(3,32)}$ = 4.658, P < 0.01; left); WT^{Veh}, WT$^{TMZ\ (High\ KD)}$, 5×FAD^{Veh}, and 5×FAD$^{TMZ\ (High\ KD)}$ mice ($F_{(3,33)}$ = 6.041, P < 0.01; middle); and WT$^{LV\text{-}GFP}$, WT$^{LV\text{-}dnWnt\ (High\ KD)}$, 5×FAD$^{LV\text{-}GFP}$, and 5×FAD$^{LV\text{-}dnWnt\ (High\ KD)}$ mice ($F_{(3,33)}$ = 5.476, P < 0.01; right). (**C**) Hippocampal PSD95 levels (left: $F_{(3,32)}$ = 5.753, P < 0.01; middle: $F_{(3,33)}$ = 3.983, P < 0.05; right: $F_{(3,33)}$ = 4.639, P < 0.01). (**D**) Representative Golgi stains in 5×FAD^{Sham} (upper) and 5×FAD^{IR} (lower) mice, and quantification of total dendritic length and neuron branch number in outer granule cell layer of 5×FAD^{Sham} (n = 3) and 5×FAD^{IR} mice (n = 4). Scale bar: 100 μm. (**E** and **F**) Mean number of errors in training days of RAM task. In (E), left: WT^{Sham}, WT^{IR}, 5×FAD^{Sham}, and 5×FAD^{IR} mice; middle: WT^{veh}, WT^{TMZ}, 5×FAD^{veh}, and 5×FAD^{TMZ} mice; right: WT$^{LV\text{-}GFP}$, WT$^{LV\text{-}dnWnt}$, 5×FAD$^{LV\text{-}GFP}$, and 5×FAD$^{LV\text{-}dnWnt}$ mice. In (F), left: 5×FAD^{Veh}, 5×FAD$^{TMZ\ (Mod\ KD)}$, and 5×FAD$^{TMZ\ (High\ KD)}$ mice; right: 5×FAD$^{LV\text{-}GFP}$, 5×FAD$^{LV\text{-}dnWnt\ (Mod\ KD)}$, and 5×FAD$^{LV\text{-}dnWnt\ (High\ KD)}$ mice.

the structural integrity of the DG specifically under pathological conditions of AD but not under normal physiological conditions.

Enzyme-linked immunosorbent assay (ELISA) and immunoblot assays revealed that the levels of postsynaptic density 95 (PSD95) and synapse-associated protein 97 (SAP97) in the hippocampal homogenates of 5×FAD$^{\text{-AHN (High KD)}}$ mice were reduced compared to 5×FAD$^{\text{CTL}}$ mice at 5 months of age (Fig. 4C and fig. S5). Meanwhile, AHN reduction did not change the levels of PSD95 and SAP97 in WT mice. These results demonstrate that suppressing AHN resulted in considerable loss of synaptic proteins PSD95 and SAP97 in the hippocampus of 5×FAD mice, likely due to DG neuronal cell death.

Golgi staining showed that the morphology of granule neurons in the outer granule cell layer, generated during early development, of 5×FAD$^{\text{IR}}$ mice did not differ from that of neurons of 5×FAD$^{\text{Sham}}$ mice (Fig. 4D and fig. S6). Furthermore, Casp3$^+$ cell number did not increase in 5×FAD$^{\text{-AHN (Mod KD)}}$ mice (Fig. 3F). These results suggest that the increase in Casp3$^+$ cells and loss of granule neurons and synaptic markers in 5×FAD$^{\text{-AHN (High KD)}}$ mice can be attributed mostly to AHN loss, rather than the direct impact of IR, TMZ, or LV-dnWnt on the existing neurons in the more susceptible brain environment of 5×FAD mice.

Five-month-old 5×FAD mice in which AHN was suppressed much later (at 4 to 4.5 months), and which had AHN levels similar to those of untreated 5×FAD mice before AHN was suppressed, exhibited only negligible Casp3$^+$ cell numbers (fig. S7A). We also could not detect any significant Casp3$^+$ cell numbers in 3-month-old 5×FAD mice in which AHN was ablated starting at 1.5 to 2 months old (fig. S7B). These results indicate

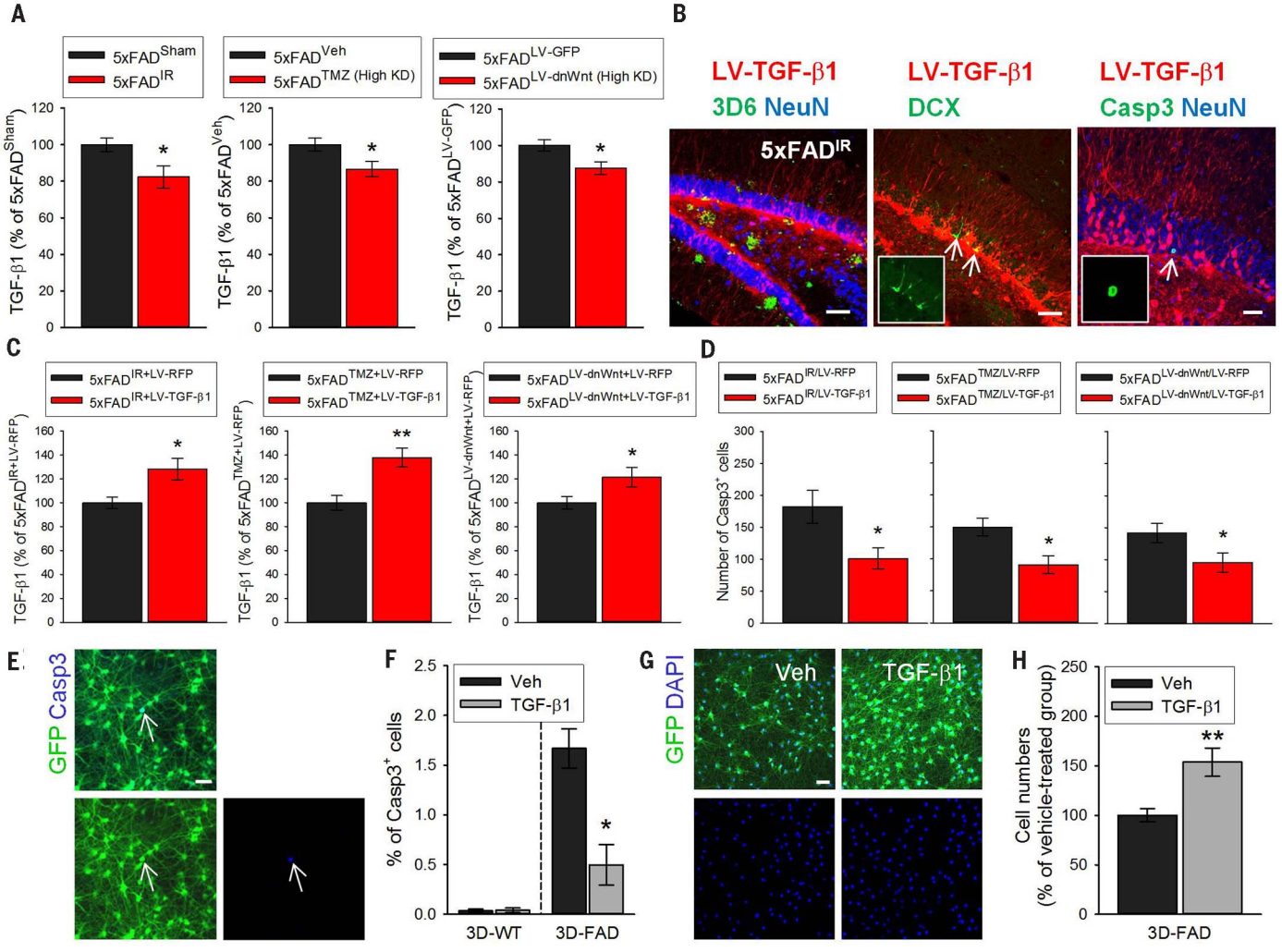


Fig. 5. Ablating AHN in male 5×FAD mice reduces hippocampal levels of TGF-β1, a protective cytokine. (**A**) Hippocampal TGF-β1 levels in male 5×FAD$^{\text{Sham}}$ and 5×FAD$^{\text{IR}}$ (left), 5×FAD$^{\text{Veh}}$ and 5×FAD$^{\text{TMZ (High KD)}}$ (middle), and 5×FAD$^{\text{LV-GFP}}$ and 5×FAD$^{\text{LV-dnWnt (High KD)}}$ mice (right). Levels are shown as percent of 5×FAD control group in each treatment. (**B**) Photomicrographs of 3D6$^+$ Aβ plaques and NeuN$^+$ cells (left), DCX$^+$ cells (middle), and Casp3$^+$ cell and NeuN$^+$ cells (right) in the transduced DG of 5×FAD$^{\text{IR}}$ mice by LV-TGF-β1. Insets represent digitally magnified images of the DCX$^+$ cells (arrows, middle) or Casp3$^+$ cell (arrow, right). Scale bars: 50 μm (left, middle); 20 μm (right). (**C**) Hippocampal TGF-β1 levels in 5×FAD$^{\text{IR/LV-RFP}}$ (n = 8) and 5×FAD$^{\text{IR/LV-TGF-β1}}$ (n = 8) (left), 5×FAD$^{\text{TMZ/LV-RFP}}$ (n = 9) and 5×FAD$^{\text{TMZ/LV-TGF-β1}}$ (n = 12) (middle), and 5×FAD$^{\text{LV-dnWnt/LV-RFP}}$ (n = 8) and 5×FAD$^{\text{LV-dnWnt/LV-TGF-β1}}$ (n = 9) (right) mice. Levels are shown as percent of 5×FAD control group in each treatment. 5×FAD$^{\text{TMZ/LV-RFP}}$ and 5×FAD$^{\text{LV-dnWnt/LV-RFP}}$ mice with moderate AHN knockdown were not included. (**D**) Quantification of Casp3$^+$ cells in 5×FAD$^{\text{IR/LV-RFP}}$ and 5×FAD$^{\text{IR/LV-TGF-β1}}$ (left), 5×FAD$^{\text{TMZ/LV-RFP}}$ and 5×FAD$^{\text{TMZ/LV-TGF-β1}}$ (middle), and 5×FAD$^{\text{LV-dnWnt/LV-RFP}}$ and 5×FAD$^{\text{LV-dnWnt/LV-TGF-β1}}$ (right) mice. (**E**) Representative images of Casp3$^+$ cells (arrows) in GFP-labeled 3D-FAD cell cultures. Scale bar: 50 μm. (**F**) Quantification of Casp3$^+$ cells in 3D-FAD cell cultures treated with vehicle (veh) or TGF-β1 (10 ng/ml). n = 3 per group. (**G**) Representative images of DAPI$^+$ cells (blue) in ReN-FAD cell cultures treated with vehicle or TGF-β1. Scale bar: 50 μm. (**H**) Number of cells survived in 3D-FAD cultures treated with vehicle or TGF-β1. n = 3 per group.

that adult-generated neurons at a relatively early stage of AD (6 weeks to 4 months old) are critical for later maintaining the survival and stability of granule neurons (at 5 months and older). Thus, AHN impairment at a very early disease stage appears to initiate cell death in later stages, i.e., when the neuronal milieu becomes more hostile [e.g., elevated levels of Aβ pathology, interleukin-1β (IL-1β), tumor necrosis factor α (TNFα), and keratinocyte-derived chemokine (KC/GRO); reduced levels of brain-derived neurotrophic factor (BDNF); figs. S2 and S3]. Blocking AHN did not affect the levels of Aβ deposition, numbers of $Iba1^+$ microglia, or $GFAP^+$ astrocytes in the hippocampus of 5-month-old 5×FAD mice (fig. S8), suggesting that AHN in 5×FAD mice had no effect on Aβ pathology or gliosis.

Ablating AHN exacerbated cognitive impairment in 5×FAD mice

We next investigated whether cognitive dysfunction progressed more rapidly in the absence of AHN in 5×FAD mice. Untreated 5×FAD mice showed significant impairment in pattern separation at 5 months compared to age-matched WT mice, but not at 3 months of age (fig. S1). Therefore, the impact of AHN loss on pattern separation memory was studied in WT^{-AHN} and $5×FAD^{-AHN}$ mice at 3 months of age. In the DNMP task, both 3-month-old WT^{-AHN} and $5×FAD^{-AHN}$ mice exhibited equally poor performances; no further impairment was observed in the $5×FAD^{-AHN}$ mice compared to WT^{-AHN} mice (fig. S9). These results suggest that adult-generated neurons are required for pattern separation both in WT and 5×FAD mice. Thus, decreased AHN in AD is sufficient to cause impairment of this memory type, and other AD pathological factors are not required for this behavioral change.

In the RAM task, 5-month-old WT^{-AHN} mice did not perform differently from WT^{CTL} mice, indicating that WT mice lacking AHN could still perform this task at a normal rate. However, ablating AHN in 5-month-old 5×FAD mice resulted in significant cognitive impairments (Fig. 4E; see table S3 for statistical analysis and group performance). However, $5×FAD^{-AHN\ (Mod\ KD)}$ mice performed comparably with $5×FAD^{CTL}$ mice (Fig. 4F; see table S4 for statistical analysis and group performance). These results suggest that a high degree of AHN knockdown, accompanied by loss of granule cells and synaptic markers, exacerbates AD-associated cognitive impairment in 5×FAD mice. Similar results were observed in the Y-maze task except that $5×FAD^{LV\text{-}dnWnt\ (Mod\ KD)}$ mice also showed impaired memory compared to $5×FAD^{LV\text{-}GFP}$ mice (fig. S10).

Five-month-old 5×FAD mice in which AHN was ablated starting at 4 to 4.5 months of age performed comparably to $5×FAD^{CTL}$ mice in the RAM task (fig. S11). No impairment was observed in 3-month-old $5×FAD^{-AHN}$ mice in which AHN was suppressed beginning at 1.5 to 2 months of age compared to those with AHN (fig. S12). These results suggest that loss of AHN, alone, is not sufficient to cause global hippocampal cognitive deficits in 5×FAD mice, and that loss of mature granule neurons (owing to AHN loss at a much earlier age) is minimally required to result in global cognitive deficits later in life (beginning at 5 months of age).

Ablating AHN in male 5×FAD mice reduces hippocampal levels of TGF-β1

The mechanisms by which the lack of AHN contributes to cell death in mature neural populations in 5×FAD mice are likely diverse with multiple pathways, and future efforts will be necessary to elucidate them. Among the cytokines we measured, transforming growth factor–β1 (TGF-β1) levels were significantly reduced in the hippocampal tissue homogenates of $5×FAD^{-AHN\ (High\ KD)}$ mice compared to $5×FAD^{CTL}$ mice (Fig. 5A and table S5). TGF-β1 levels were not reduced in $WT^{-AHN\ (High\ KD)}$ mice compared to WT^{CTL} mice (fig. S13), suggesting that AHN loss, accompanied by loss of granule cells and synaptic markers, results in the reduced TGF-β1 levels. TGF-β1 is an anti-inflammatory cytokine that increases rapidly after injury and with age (*22*). Neurons and multipotent NPCs produce TGF-β1 (*23*, *24*), and TGF-β1 inhibits Casp3 (*25*). We asked whether TGF-β1 has protective effects against cell death in AD, i.e., whether reduced TGF-β1 levels in $5×FAD^{-AHN}$ contribute to cell death. For this purpose, we increased TGF-β1 in $5×FAD^{-AHN}$ mice by injecting lentivirus expressing TGF-β1 (LV-TGF-β1; $5×FAD^{IR/LV\text{-}TGF\text{-}β1}$, $5×FAD^{TMZ/LV\text{-}TGF\text{-}β1}$, and $5×FAD^{LV\text{-}dnWnt/LV\text{-}TGF\text{-}β1}$) and quantified $Casp3^+$ cells (Fig. 5, B to D). These mice experienced increased hippocampal TGF-β1 levels and decreased numbers of $Casp3^+$ cells in the DG compared to 5×FAD mice without AHN that were injected with control LV-RFP ($5×FAD^{IR/LV\text{-}RFP}$, $5×FAD^{TMZ/LV\text{-}RFP}$, and $5×FAD^{LV\text{-}dnWnt/LV\text{-}RFP}$). However, $Casp3^+$ cells were still evident in the $5×FAD^{-AHN}$ mice injected with LV-TGF-β1, and increasing TGF-β1 levels was not sufficient to rescue cell death fully in 5×FAD mice without AHN. Our results suggest that TGF-β1 has protective effects against cell death in 5×FAD mice and that the reduced TGF-β1 levels in 5×FAD mice without AHN, at least in part, contribute to granule cell death. Other mechanisms that maintain existing neuronal stability in AD likely account for the fraction of cell death that was not rescued by TGF-β1. Increasing TGF-β1 in $5×FAD^{-AHN}$ mice did not change the degree of AHN (fig. S14), although TGF-β1 has been shown to regulate neurogenesis (*26*–*29*).

Protective effects of TGF-β1 against cell death were confirmed in three-dimensional (3D) cultures of differentiated human NPCs expressing familial AD-related mutations (3D-FAD cells) (*30*). Our 3D-FAD cells showed an increased number of $Casp3^+$ cells and endogenous neuronal loss, as indicated by cell counting, compared to control WT cells (3D-WT cells) beginning at 5 weeks of differentiation (fig. S15). Six-week differentiated 3D-FAD cells were treated with TGF-β1 for up to 2 weeks, after which cell viability was assessed. $Casp3^+$ cell number was significantly reduced in 3D-FAD cells treated with TGF-β1 (10 ng/ml; Fig. 5, E and F). Treatment of TGF-β1 significantly increased cell survival in 3D-FAD cells, as indicated by cell counting (Fig. 5, G and H). Furthermore, a lactate dehydrogenase (LDH) assay showed that treatment with TGF-β1 resulted in decreased LDH release (fig. S16A). Protective effects of TGF-β1 were also shown by a CellTiter-Glo luminescent assay, in which cell viability was increased with TGF-β1 treatment (fig. S16B).

Endogenous levels of TGF-β1 were measured in 0.5-, 5-, and 8-week differentiated 3D-WT and 3D-FAD cultures (fig. S17A). Considering that our 3D cultures are composed of mostly neurons, with astrocyte presence, and do not contain microglia, the endogenous TGF-β1 detected in our 3D culture models was secreted from neurons and/or astrocytes. At 0.5 week of differentiation, TGF-β1 levels were higher in the 3D-FAD cultures compared to 3D-WT. Whereas TGF-β1 levels in the 3D-WT did not differ significantly at the three time points, in the 3D-FAD culture, levels of TGF-β1 were significantly reduced in the 5- and 8-week differentiated cultures, where cell death occurred, compared to the 0.5-week differentiated cultures. When normalized by actin signal density, TGF-β1 levels in the 5- and 8-week differentiated 3D-FAD cultures did not differ from those in age-matched 3D-WT cultures (fig. S17B), suggesting that the decrease in TGF-β1 observed in 3D-FAD cultures compared to 3D-WT cultures was due to cell loss. However, we also observed that TGF-β1 levels normalized by actin signal were significantly lower in the 8-week differentiated 3D-FAD cultures compared to 5 week-differentiated 3D-FAD cultures. Therefore, we cannot exclude the possibility that the reduced levels observed at the 8-week time point were also attributable to specific signaling inhibitions.

Effects of ablating AHN in female 5×FAD mice

AHN loss in female 5×FAD mice was accompanied by an increased number of $Casp3^+$ cells in the DG, reduced hippocampal levels of PSD95, and exacerbated cognitive dysfunction (Fig. 6, A to E, and fig. S18; see fig. S18A for number of animals in each experimental group and group arrangement explanations). However, these mice did not show reduced hippocampal levels of TGF-β1 (Fig. 6F). Thus, we measured endogenous levels of TGF-β1 in the hippocampus of untreated male and female WT and 5×FAD mice at different ages (Fig. 6G). Untreated female 5×FAD mice showed significantly increased TGF-β1 levels as compared to WT mice and increased, although not statistically significant, TGF-β1 levels as compared to male 5×FAD mice at 5 months of age. The relatively increased levels of TGF-β1 in female 5×FAD mice might mask the ability to observe a small reduction of TGF-β1, as observed in male 5×FAD mice without AHN. Therefore, blocking AHN might result in reduction of TGF-β1 in 5×FAD mice at an age when endogenous TGF-β1 levels are not yet increased as compared to WT mice. AHN loss still increased $Casp3^+$ cell numbers in female 5×FAD mice, suggesting that reduction of TGF-β1 is most likely not the only mechanism underlying cell death in 5×FAD mice lacking AHN.

Exercise increased levels of hippocampal BDNF, PSD95, SYP, IL-6, and FNDC5

We addressed the mechanism by which the combination of exercise and increased AHN, but not increased AHN alone, improved cognition in the 5×FAD mice. Exercise has been shown to increase neurotrophins, growth factors, and synaptic markers and to reduce neuroinflammation (*31*–*35*). Exercise induces fibronectin type III domain–containing protein–5 (FNDC5) that regulates hippocampal BDNF expression in mice (*36*). We first measured 18 molecules in the hippocampus of untreated sedentary and exercised 5×FAD mice and found that exercise increased the levels of hippocampal BDNF, PSD95, synaptophysin (SYP), interleukin-6 (IL-6), and FNDC5 (table S6). We next measured the levels of these molecules in the 5×FAD^{CTL}, 5×FAD^{ProAHN}, 5×FAD$^{+AHN(RUN)}$, and 5×FAD$^{\psi AHN(RUN)}$ mice. Exercise again led to increased levels of hippocampal BDNF, PSD95, SYP, IL-6, and only in male mice, FNDC5, whereas activating AHN alone did not (Table 1 and table S7). Considering that AD patients have reduced levels of BDNF, PSD95, and SYP (*37*–*39*) and that these proteins are important regulators of synaptic plasticity, the increase in these protein levels with exercise combined with increased AHN may contribute to an improved behavioral outcome for patients, as observed in the 5×FAD$^{+AHN(RUN)}$ mice. IL-6, for which levels are selectively elevated in the hippocampus following exercise (*40*), has also been reported to benefit cognition and regulate neurogenesis (*41*–*45*).

AHN activation combined with increased BDNF levels genetically ameliorates cognitive function in 5×FAD mice

To explore the mechanisms underlying cognitive improvement after combined exercise and increased AHN, we directly increased BDNF, IL-6, or FNDC5 in the hippocampus of 5×FAD^{ProAHN} mice. We investigated whether increasing levels of these factors in conjunction with increased AHN would result in cognitive improvements mimicking those observed in the 5×FAD$^{+AHN(RUN)}$ mice. To increase BDNF in the 5×FAD^{ProAHN} mice, lentivirus expressing BDNF (LV-BDNF) was injected into the hippocampus at 3 months of age when the mice received LV-Wnt3 (5×FAD$^{ProAHN/LV-BDNF}$; Fig. 7A). Lentivirus expressing red fluorescent protein (LV-RFP) was injected into additional 5×FAD^{ProAHN}

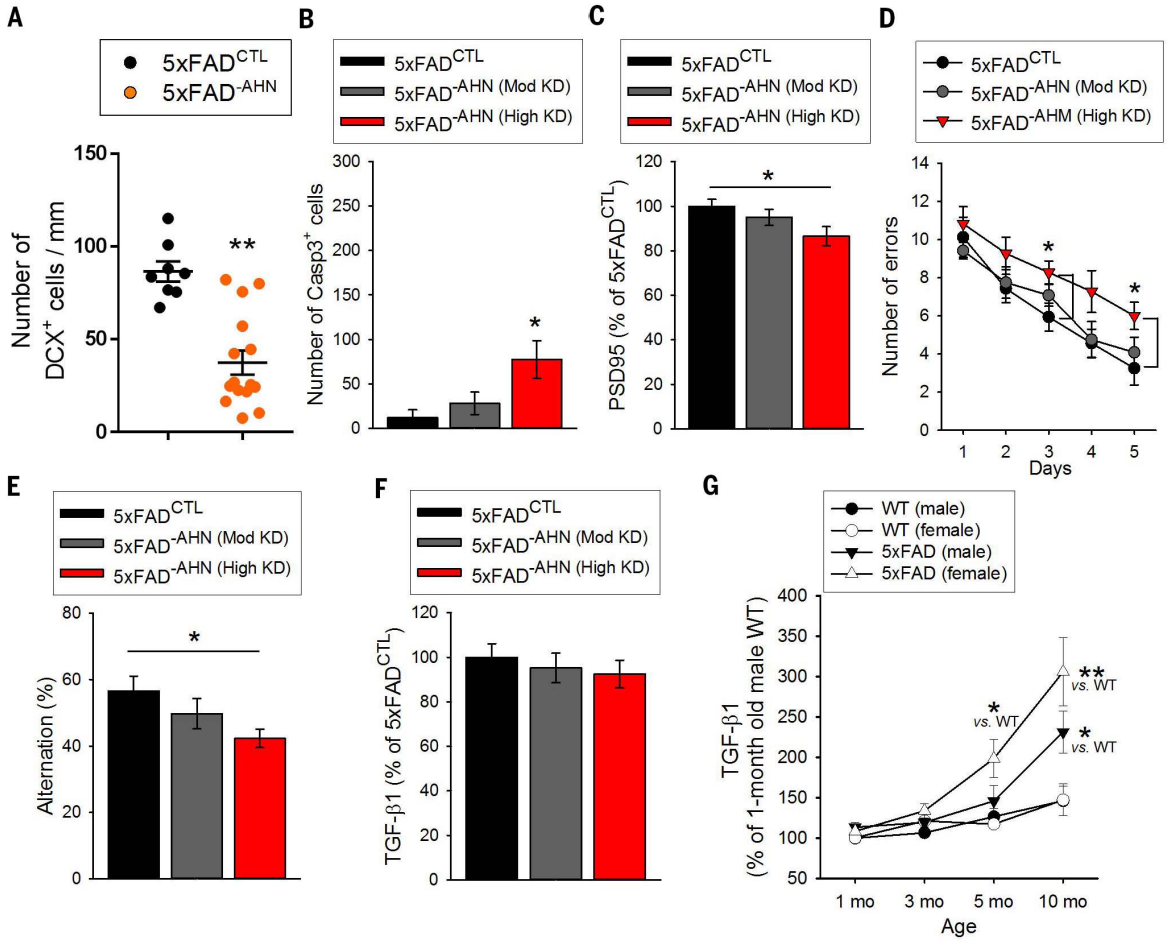


Fig. 6. Effects of ablating AHN in female 5×FAD mice. (**A**) Quantification of DCX$^+$ cells in female 5×FAD^{CTL} and 5×FAD^{-AHN} mice. (**B**) Quantification of Casp3$^+$ cells in female 5×FAD^{CTL}, 5×FAD$^{-AHN\ (Mod\ KD)}$, and 5×FAD$^{-AHN\ (High\ KD)}$ mice ($F_{(2,20)}$ = 4.803, P < 0.05). (**C**) Levels of hippocampal PSD95 ($F_{(2,20)}$ = 3.499, P < 0.05). Levels are shown as percent of 5×FAD^{CTL} group. (**D**) Mean error number in RAM task training trials. Two-way ANOVA with repeated measures revealed significant effects for days ($F_{(4,80)}$ = 24.19, P < 0.01) and groups ($F_{(2,20)}$ = 5.982, P < 0.01) but not interaction ($F_{(8,80)}$ = 0.4305, P = 0.8994). Analysis of error number on each day by Fisher's LSD post hoc tests revealed that 5×FAD$^{-AHN\ (High\ KD)}$ differed significantly from 5×FAD^{CTL} mice on days 3 and 5 (day 3, $F_{(2,20)}$ = 3.495, P < 0.05; day 5, $F_{(2,20)}$ = 3.419, P = 0.0428). (**E**) Spontaneous alternation in Y-maze ($F_{(2,20)}$ = 3.747, P < 0.05). Total arm entries comparable among groups (fig. S18G). (**F**) Hippocampal TGF-β1 levels. Levels are shown as percent of 5×FAD^{CTL} group. (**G**) Changes in TGF-β1 levels in the hippocampal homogenates of untreated male and female WT and 5×FAD mice with age (n = 7 per group). In 5-month-olds, *P < 0.05 between female 5×FAD and WT mice. In 10-month-olds, *P < 0.05 between male 5×FAD and WT mice; **P < 0.01 between female 5×FAD and WT mice.

and 5×FAD$^{+AHN(RUN)}$ mice (5×FAD$^{ProAHN/LV-RFP}$ and 5×FAD$^{+AHN(RUN)/LV-RFP}$ mice, respectively). LV-BDNF injection increased hippocampal BDNF levels in both 6-month-old male and female 5×FAD^{ProAHN} mice (Fig. 7B; see table S8 for statistical analysis of Fig. 7 experiments). Although BDNF has been shown to regulate AHN positively (*46*), LV-BDNF injection did not further increase the number of DCX$^+$ neurons in 5×FAD mice treated with P7C3 and LV-Wnt3, and the number of DCX$^+$ neurons was comparable in all groups (Fig. 7C). LV-BDNF injection did not affect Aβ plaque levels (Fig. 7D).

In the Y-maze, both male and female 5×FAD$^{+AHN(RUN)/LV-RFP}$ mice showed improved memory compared to the 5×FAD$^{ProAHN/LV-RFP}$ mice of both genders (Fig. 7E and fig. S19). 5×FAD$^{ProAHN/LV-BDNF}$ mice also exhibited similar memory improvement. In male cohorts, during the RAM task training trials, 5×FAD$^{+AHN(RUN)/LV-RFP}$ mice showed improved memory compared to 5×FAD$^{ProAHN/LV-RFP}$ mice (Fig. 7F, left graph). Although 5×FAD$^{ProAHN/LV-BDNF}$ mice did not show significant improvement, their performance did not differ from that of either 5×FAD$^{ProAHN/LV-RFP}$ or 5×FAD$^{+AHN(RUN)/LV-RFP}$ mice. However, in the memory retention test, 5×FAD$^{ProAHN/LV-BDNF}$ mice showed memory improvement compared to 5×FAD$^{ProAHN/LV-RFP}$ mice and behaved similarly to 5×FAD$^{+AHN(RUN)/LV-RFP}$ mice (Fig. 7F, right graph). Increasing AHN alone by P7C3 and LV-Wnt3 did not improve pattern separation memory in female 5×FAD mice (Fig. 2A, right graph). However, increasing AHN in conjunction with BDNF in female 5×FAD mice (female 5×FAD$^{ProAHN/LV-BDNF}$ mice) produced significantly improved pattern separation memory (Fig. 7G).

Increasing BDNF alone, in the absence of promoting AHN by P7C3 and LV-Wnt3, failed to increase AHN or improve memory in 5×FAD mice (Fig. 7, H and I, and fig. S20). These results suggest that increasing hippocampal BDNF, in combination with AHN activation, is sufficient to mimic the beneficial effects of exercise on cognition in 5×FAD mice, even in the continued presence of Aβ plaques.

AHN activation, combined with increased BDNF levels, pharmacologically ameliorates cognitive function in 5×FAD mice

We next explored whether a late-stage increase in BDNF pharmacologically (AICAR), with increased AHN by P7C3 and LV-Wnt3, could also provide an effective therapeutic strategy comparable to that of sustained exercise in 5×FAD mice. For this purpose, we used the AMP-activated protein kinase agonist 5-aminoimidazole-4-carboxamide riboside (AICAR), which increases BDNF in mice (*47*). Male and female 5×FAD^{ProAHN} mice were injected with AICAR or saline every other day for 2 weeks starting at 5.5 months of age; they were sacrificed at 6 months of age (5×FAD$^{ProAHN/AICAR}$ and 5×FAD$^{ProAHN/Veh}$ mice, respectively). Hippocampal BDNF levels were increased in the male and female 5×FAD$^{ProAHN/AICAR}$ mice compared to 5×FAD$^{ProAHN/Veh}$ mice, and their levels were not different from those of 5×FAD$^{+AHN(RUN)}$ mice injected with saline (5×FAD$^{+AHN(RUN)/Veh}$; Fig. 7J). The number of DCX$^+$ neurons was comparable among all groups (fig. S21A).

Male mice were tested in the Y-maze and RAM tasks, and female mice were tested in the Y-maze and DNMP tasks. In the Y-maze, 5×FAD$^{ProAHN/AICAR}$ mice showed improved memory compared to the 5×FAD$^{ProAHN/LV-RFP}$ mice in each respective gender, and their performance did not differ from that of 5×FAD$^{+AHN(RUN)/Veh}$ mice (Fig. 7K and fig. S21B). Although male 5×FAD$^{ProAHN/AICAR}$ mice did not perform better than 5×FAD$^{ProAHN/Veh}$ mice in RAM task training trials (fig. S21C), they showed better cognition in the retention test of the task, performing similarly to 5×FAD$^{+AHN(RUN)/Veh}$ mice (Fig. 7L). Female 5×FAD$^{ProAHN/AICAR}$ mice showed improved pattern separation in the DNMP task, again performing similarly to 5×FAD$^{+AHN(RUN)/Veh}$ mice (Fig. 7M).

It has been reported that AICAR alone increases AHN and cognition in WT mice (*47*, *48*). However, AICAR alone failed to increase AHN in the 5×FAD mice, although it increased hippocampal BDNF levels (fig. S22A). No effect of AICAR was observed in the cognition tasks that we employed in 5×FAD mice with no AHN activation (fig. S22B-S22D). These results suggest that a late-stage increase in BDNF by AICAR, in the presence of promoted AHN, could mimic the beneficial effects of exercise on cognition in 5×FAD mice.

Future studies will be needed to explore the mechanisms by which increasing levels of BDNF and AHN combine to mimic the benefits of exercise on cognition in AD. LV-BDNF injection increased levels of hippocampal PSD95, whereas AICAR did not (table S9). Neither LV-BDNF nor AICAR increased IL-6 levels (table S9). These results suggest that the beneficial effects of increasing both AHN and BDNF on cognition could be independent of PSD95 or IL-6. Directly increasing levels of hippocampal IL-6 or FNDC5 via lentivirus failed to improve cognition or increase BDNF in 5×FAD^{ProAHN} mice. However, it is possible that only the secreted form of peripherally expressed FNDC5 is responsible for exercise-induced hippocampal BDNF, as peripheral delivery of FNDC5 induces the expression of hippocampal BDNF (*36*). There may be other possible effects of AICAR aside from increasing BDNF levels. TGF-β1 levels were not changed by increasing AHN with or without increased BDNF, increasing BDNF only, or by exercise in our study (table S10).

Discussion

In the present study, we did not explore the effects of AHN manipulation on tau phosphorylation levels because basal tau phosphorylation is already present in adult-generated neurons at DCX$^+$ immature neuronal stages [(*49*); see fig. S23]. Future studies are necessary to investigate whether adult-generated neurons affect tauopathy in the human AD brain. The effects of physical exercise on cognition in patients with dementia are inconclusive (*50*–*52*). In our study, 5×FAD$^{ψAHN(RUN)}$ mice did not show improved cognitive function, suggesting that increased AHN is required to mediate the beneficial effects of exercise in 5×FAD mice. Individuals who

Table 1. Levels of hippocampal BDNF, PSD95, SYP, IL-6, and FNDC5 of male 5×FAD^{CTL}, 5×FAD^{ProAHN}, 5×FAD$^{+AHN(RUN)}$, and 5×FAD$^{ψAHN(RUN)}$ mice. Animal numbers are in parentheses (results for female mice are in table S7). $F_{(3,31)} = 11.48$, $P < 0.01$ (BDNF); $F_{(3,41)} = 6.279$, $P < 0.01$ (PSD95); $F_{(3,41)} = 6.083$, $P < 0.01$ (SYP); $F_{(3,41)} = 6.047$, $P < 0.01$ (IL-6); $F_{(3,31)} = 5.974$, $P < 0.01$ (FNDC5). *$P < 0.05$; **$P < 0.01$ compared to 5×FAD^{CTL} and 5×FAD^{ProAHN} mice.

	5×FAD^{CTL}	5×FAD^{ProAHN}	5×FAD$^{+AHN(RUN)}$	5×FAD$^{ψAHN(RUN)}$
BDNF	100.00 ± 5.72 (8)	91.32 ± 3.80 (10)	149.74 ± 10.65** (8)	145.29 ± 13.09** (9)
PSD95	100.00 ± 7.07 (10)	103.19 ± 5.64 (15)	129.39 ± 6.47* (12)	128.23 ± 4.41* (8)
SYP	100.00 ± 6.02 (10)	102.03 ± 4.21 (15)	125.59 ± 7.04* (12)	125.02 ± 4.39* (8)
IL-6	100.00 ± 2.58 (10)	105.44 ± 3.86 (15)	128.89 ± 9.80* (12)	129.52 ± 6.74* (8)
FNDC5	100.00 ± 6.54 (8)	100.16 ± 7.49 (10)	131.78 ± 6.26* (8)	127.41 ± 7.03* (9)

remain cognitively intact despite showing neuropathological AD features have increased numbers of neural stem cells compared to AD subjects (*53*). Therefore, future studies are required regarding whether patients with dementia who experienced physical exercise have increased AHN enough to have beneficial effects of exercise on cognition.

In summary, inducing AHN alone, e.g., pharmacologically and genetically, conferred only minimal to no benefit in 5×FAD mice. However, exercise-induced AHN improved cognition along with reduced Aβ load and increased levels of BDNF, IL-6, FNDC5, and synaptic markers. AHN activation was also required for exercise-induced

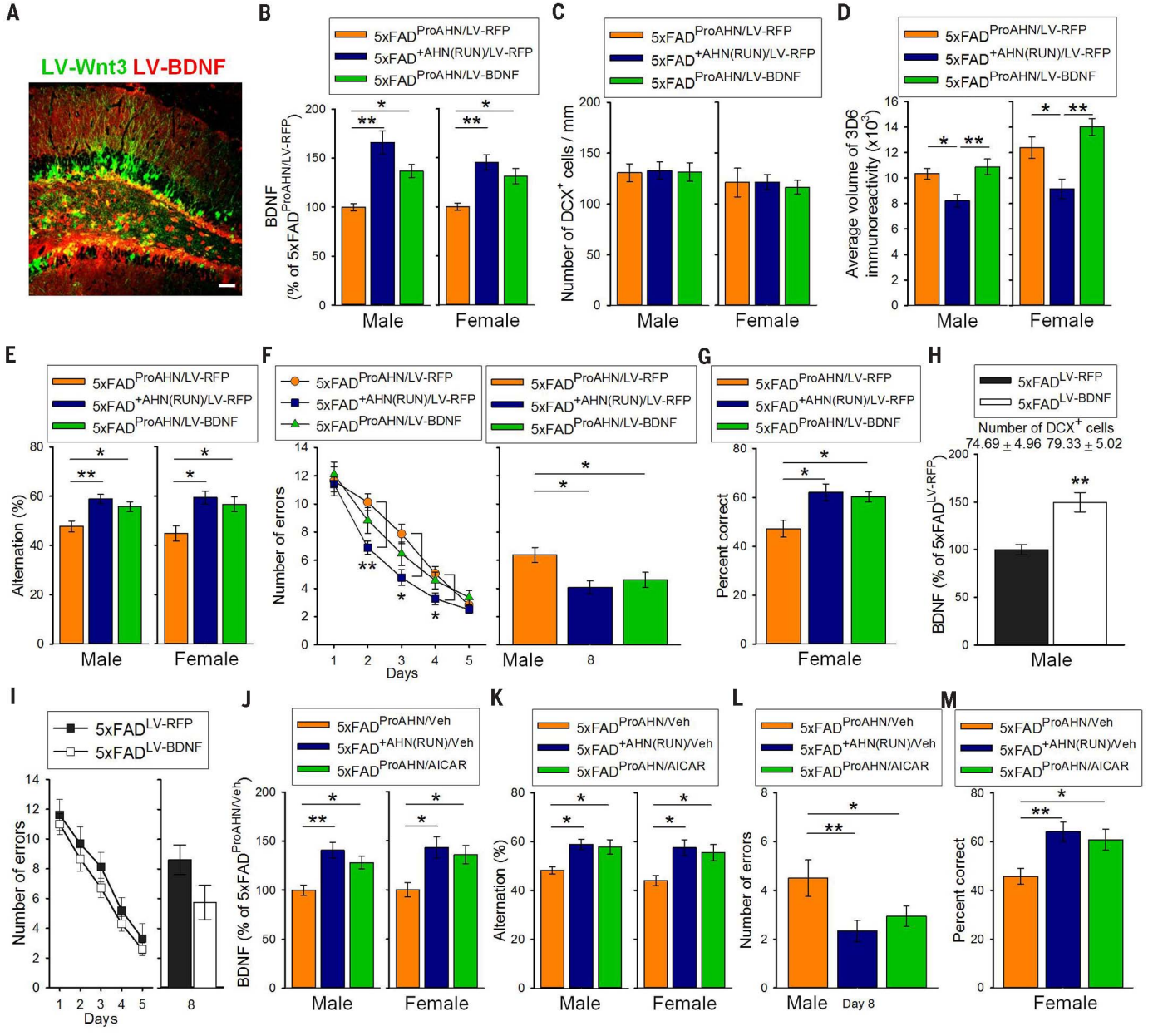


Fig. 7. AHN activation combined with increased BDNF levels ameliorates cognitive function in 5×FAD mice. (**A**) Photomicrograph of lentiviral expression of LV-Wnt3 and LV-BDNF in the DG of 5×FAD$^{ProAHN/LV\text{-}BDNF}$ mice. Scale bar: 50 μm. (**B**) Hippocampal BDNF levels in 5×FAD$^{ProAHN/LV\text{-}RFP}$, 5×FAD$^{+AHN(RUN)/LV\text{-}RFP}$, and 5×FAD$^{ProAHN/LV\text{-}BDNF}$ mice. (**C**) DCX$^+$ cell quantification. (**D**) Quantitative analysis of Aβ burden volume (mean voxel count ± SEM). (**E**) Spontaneous alternation behavior in Y-maze task. Total arm entries comparable among groups (fig. S19). (**F**) Left: mean error number for each group (male 5×FAD$^{ProAHN/LV\text{-}RFP}$, 5×FAD$^{+AHN(RUN)/LV\text{-}RFP}$, and 5×FAD$^{ProAHN/LV\text{-}BDNF}$) in RAM task training days. Right: mean error number in memory retention trial. (**G**) Quantification of percent correct during choice phase of DNMP task among female 5×FAD$^{ProAHN/LV\text{-}RFP}$, 5×FAD$^{+AHN(RUN)/LV\text{-}RFP}$, and 5×FAD$^{ProAHN/LV\text{-}BDNF}$ mice. (**H**) Hippocampal BDNF levels in male 5×FAD$^{LV\text{-}RFP}$ (n = 8) and 5×FAD$^{LV\text{-}BDNF}$ (n = 12) mice. DCX$^+$ cell number per mm listed above graph. (**I**) Mean error number in RAM task training trials (left) and mean error number in memory retention trial (right). (**J**) Hippocampal BDNF levels in 5×FAD$^{ProAHN/Veh}$, 5×FAD$^{+AHN(RUN)/Veh}$, and 5×FAD$^{ProAHN/AICAR}$ mice. (**K**) Spontaneous alternation behavior in Y-maze task. Total arm entries comparable among groups (fig. S21B). (**L**) Mean error number in memory retention trial of RAM task. (**M**) Quantification of percent correct during choice phase of DNMP task among female 5×FAD$^{ProAHN/Veh}$, 5×FAD$^{+AHN(RUN)/Veh}$, and 5×FAD$^{ProAHN/AICAR}$ mice.

improvement in memory. These data suggest that promoting AHN can ameliorate AD pathology and cognitive deficits but only in the presence of a healthier, improved local brain environment, e.g., stimulated by physical exercise. Increasing AHN alone (genetically and pharmacologically) combined with overexpression of BDNF was able to mimic exercise-induced improvements in cognition, without reducing Aβ burden. We also found that deficits in AHN in very early stages of life can exacerbate neuronal vulnerability in AD later in life, leading to cognitive impairment and increased neuronal loss, due at least in part to reduced TGF-β1. These results suggest that adult-born neurons generated very early in life are critical for maintaining hippocampal neuronal populations in the abnormal and hostile brain environment created by AD later in life. Thus, AHN impairment may be a primary event that later mediates other aspects of AD pathogenesis. Future attempts to create pharmacological mimetics of the benefits of exercise on both increased AHN and the health of the local neuronal environment, especially involving increased BDNF levels, may someday provide an effective means for improving cognition in AD. Moreover, increasing neurogenesis in the earliest stages of AD pathogenesis may protect against neuronal cell death later in the disease, providing a potentially powerful disease-modifying treatment strategy for AD.

Materials and Methods

Animals

We used 5×FAD APP/PS1 doubly transgenic mice that co-overexpress and co-inherit FAD mutant forms of human amyloid precursor protein (APP) (the Swedish mutation: K670N, M671L; the Florida mutation: I716V; the London mutation: V717I) and presenilin 1 (PS1) (M146L; L286V) transgenes under transcriptional control of the neuron-specific mouse Thy-1 promoter (Tg6799 line). 5×FAD lines (B6/SJL genetic background) were purchased from Jackson Laboratory and were maintained by crossing heterozygous transgenic mice with B6/SJL F1 breeders. All 5×FAD transgenic mice were heterozygotes with respect to the transgene. Animal experiments were conducted in accordance with institutional and NIH guidelines.

Observed benefits of employing the 5×FAD mouse model

Most of the current AD transgenic mouse models are generated by the overexpression of mutation (s) related to familial AD (FAD). While the 5× FAD mouse model expresses rare early-onset FAD mutations, the pathological features found in the brains of 5×FAD mice are common to all forms of AD. In the current study, we examined the manner in which altered AHN affects cell survival and cognition under these pathological conditions. Hence, we believe that the results of our experiments are also relevant for generalized AD pathology (beyond that of early-onset FAD).

We chose 5×FAD mice, which show faster AD progression compared to other AD transgenic mice, based on our goal of investigating whether impaired AHN at an early stage in AD resulted in neuronal/synaptic loss and in accelerated cognitive impairment at later stages of the disease. We also aimed to investigate whether adult-generated neurons affected the pre-existing neuronal populations and whether impaired AHN exacerbated neuronal loss in AD.

5×FAD mice are one of the few AD transgenic mouse models that exhibit neurodegeneration, thus rationalizing our use of this model. In addition, performing long-term manipulations of AHN is challenging and can compromise the health of mice over time. For example, multiple irradiations cause neuro-inflammation. Long-term injections of TMZ cause side effects, including weight loss and hair loss (personal observations). Lentivirus is delivered to the brain using stereotaxic surgery, and multiple brain surgeries are not possible. Thus, studying the 5×FAD line, which undergoes more rapid AD progression compared to other AD transgenic mice, is advantageous for our study.

Generation of LV-Wnt3 and LV-dnWnt

To construct the Wnt3-Internal ribosomal entry site (IRES)-GFP vector, we cloned the cDNA of Wnt3 upstream of the IRES and GFP and inserted the bicistronic cassette in place of the GFP sequence in the pRRL.SIN.cPPT.hPGK.GFP.Wpre vector. To generate the dnWnt-IRES-GFP vector, we cloned the cDNA for dnWnt upstream of the IRES and GFP and inserted the bicistronic cassette in place of the GFP sequence in the CSC. cPPT.hCMV.GFP.Wpre vector. Concentrated lentiviral stocks were produced by calcium phosphate transfection into 293T cells. Supernatants were collected, passed through a 0.22-μm filter, and purified by ultracentrifugation. Viral stocks were stored at −80°C until use and were diluted at 0.8×10^9 to 1×10^9 transducing units/ml.

Treatment of aminopropyl carbazole (P7C3) and LV-Wnt3, and exercise setting

To increase neurogenesis, 2-month-old 5×FAD mice received 20 mg/kg of P7C3 (Sigma, St. Louis, MO; Asinex, Winston-Salem, NC) once daily for 4 or 5 days a week intraperitoneally (i.p.) over the span of 4 to 4.5 months. P7C3 was prepared for dosing by dissolving a recrystallized stock in dimethyl sulfoxide (DMSO) at 50 mg/ml. The compound was diluted to a final formulation of 3% DMSO/10% cremophor EL (Sigma, St. Louis, MO)/87.5% D5W (5% dextrose in water, pH 7.2). At the age of 3 months, these mice also received 4 intrahippocampal injections of 0.5 μl (total 2.0 μl) LV-Wnt3 viral suspension at these coordinates: AP, −2.0; ML, ±1.5-1.7; DV, −2.0 and AP, −3.0; ML, ±3.0; DV, −3.0. Injections were performed using a 5 μl Hamilton syringe with a 30-gauge needle attached to a digital stereotaxic apparatus and an infusion pump at a rate of 0.15 μl/min. After virus infusion was completed, the needle remained in place for 10 min before slow withdrawal. The mice were not injected with P7C3 both 3 days prior to and following the LV-Wnt3 injection day. Control 5×FAD mice received vehicle solution [3% DMSO/10% cremophor EL (Sigma, St. Louis, MO)/87.5% D5W (5% dextrose in water, pH 7.2) with no P7C3] and LV-GFP.

Both groups were either singly housed either in standard laboratory cages (27 cm by 11 cm by 17 cm) only or singly spent 3 hours in rat cages (45 cm by 20 cm by 20 cm) without running wheels and then were returned to their original cages for 21 hours. A cohort of control 5×FAD mice treated with vehicle and LV-GFP singly spent 3 hours in rat cages equipped with running wheels and was returned to their original cages for the remaining 21 hours, as described previously (*11*), from between the ages of 6 weeks and 2 months until the end of the experiments.

To evaluate potential changes in neurogenesis, we determined the number of DCX^+ cells in the targeted (GFP-expressing) areas. Areas infected by LVs were identified by expression of GFP.

Blocking AHN

To block AHN, 1.5- to 2-month-old WT and 5×FAD mice received IR, TMZ or LV-dnWnt and were sacrificed at 3 or 5 months of age. Irradiation was not used for mice that were sacrificed at 3 months of age because it requires ~1 month of recovery time. Additional 4- to 4.5-month-old 5×FAD mice received TMZ or LV-dnWnt and were sacrificed at 5 months of age.

Irradiation procedure

Mice were anesthetized with ketamine and xylazine, placed in a stereotaxic frame, and put into a custom-made radiation shield consisting of a chamber with 1-inch-thick lead and a 2.5-mm borehole. The shield covered the entire body but left a 2.5-mm treatment field above the hippocampus. Mice received irradiation once at the dose of 1000 cGy, administered at ~90 cGy per minute for a total irradiation time of 11 min. A Gammacell 40 Exactor (Theratronics, Ottawa, Ontario) with two cesium-139 sources was used.

TMZ injection

The DNA-alkylating agent TMZ (Merck & Co., Inc., Whitehouse Station, NJ; Sigma, St. Louis, MO) was dissolved in phosphate-buffered saline (PBS) or in DMSO followed by diluting in PBS to a concentration of 2.5 mg/ml (10% DMSO for TMZ from Sigma). TMZ was administered i.p. at a dose of 12.5 mg/kg once daily for 3 consecutive days, followed by 4 days of no injections (one cycle). After every four cycles, mice were given a 2-week rest period. Vehicle solution was the identical DMSO/PBS solution but without TMZ and was administered in volumes consistent with TMZ dosing.

LV-dnWnt injection

Stereotaxic injections of LV-dnWnt were performed as described above for LV-Wnt3 injections.

Injections of LV-TGF-β1 and LV-BDNF

LV-TGF-β1, LV-BDNF, LV-IL-6, IL-FNDC5, and control LV-RFP were purchased from ViGene Biosciences (Rockville, MD). $5\times FAD^{Sham}$, $5\times FAD^{IR}$, $5\times FAD^{Veh}$, and $5\times FAD^{TMZ}$ mice received four

intrahippocampal injections of 0.5 μl (total 2.0 μl) of LV-TGF-β1 or LV-RFP viral suspension (AP, −2.0; ML, ±1.5 to 1.7; DV, −2.0 and AP, −3.0; ML, ±3.0; DV, −3.0) at the age of 2.5 to 3 months. For injections of LV-dnWnt along with LV-TGF-β1 or LV-RFP in 5×FAD mice, 1.5- to 2-month-old mice received four intrahippocampal injections of total 2.4 to 3.0 μl of LV-dnWnt with LV-TGF-β1 or LV-RFP viral suspension. For injection of LV-BDNF, 5×FAD^{ProAHN} mice received LV-BDNF at the age of 3 months when they received LV-Wnt3: total 2.4 to 3.0 μl of LV-Wnt3 with LV-BDNF or LV-RFP viral suspension.

Treatment of 5-aminoimidazole-4-carboxamide-1-β-D-ribofuranoside (AICAR)

Mice were injected (i.p.) with AICAR (Toronto Research Chemicals Inc., Canada) dissolved in saline, 250 mg/kg per day, or with saline, every other day for 2 weeks. During the 2 weeks, P7C3 was not injected.

5-Bromo-2'-deoxyuridine (BrdU) injections

BrdU (Sigma, St. Louis, MO) was dissolved in 0.9% NaCl at a concentration of 20 mg/ml and was filtered (0.2 μm) under sterile conditions. Mice received a single i.p. injection of BrdU (50 mg/kg) daily for 3 days and were perfused 1 day after the last BrdU injection and processed for BrdU immunostaining to identify proliferating neural progenitor cells (NPCs). Parallel cohorts of animals from each group were sacrificed 4 weeks after the last BrdU injection and processed to determine the rate of NPC survival and differentiation.

Behavior tests

We performed three different behavioral tests: a delayed nonmatching to place (DNMP) task to measure pattern separation, an eight-arm radial arm maze (RAM) to measure reference memory and retention memory, and a Y-maze to measure short-term spatial (working) memory. For the DNMP and RAM tasks, mice were food restricted to 90% of their pre-experimental free-feeding weights with water available ad libitum. We confirmed that our food deprivation regime did not affect AHN in WT and 5×FAD mice. Behavior analyses were performed blindly.

DNMP task

Pattern separation-dependent memory was tested in the DNMP task in the RAM apparatus as described previously (*16*) with modifications. The testing apparatus was a rat-sized RAM. The maze consisted of eight equally spaced arms that radiated from an octagonal central platform. Arms were 15 cm wide by 75 cm long, and the walls on each arm were 2.5 cm high. On the morning before testing commenced, mice were habituated to the maze, where all arms were unblocked and all wells at the end of the arms contained several sunflower seeds. Mice were allowed to explore the maze freely during this habituation session for 30 min. Each trial of this task consisted of a sample phase and a choice phase.

During the sample phase, all arms except a start arm and the sample (rewarded) arm were blocked off. Each mouse was permitted to visit the sample arm and retrieve a food pellet reward. Mice were retrieved from the maze after either (1) spending 10 s in the sample arm after retrieving the pellet or (2) exiting the sample arm. During the choice phase, arms in the start and sample (unrewarded) locations and an additional correct (rewarded) location were open. The correct arm was distant from the sample arm by a spatial separation of two arms. In this phase, mice were tested for the ability to select, from a choice of two arms, the arm location that had not been presented in a previous sample phase. Mice that entered the correct (new, rewarded) arm were considered to have made correct choices. Mice that entered an incorrect (familiar, unrewarded) arm were allowed to self-correct. Mice went through four trials (sample plus choice phases) per day for 3 consecutive days.

RAM

The RAM was used for testing spatial reference and retention memory. To adapt to the maze and bait (pretraining), the mice experienced free movement and feeding in a RAM twice a day for 2 days. The bait was scattered in all arms and a sunflower seed in a food cup was available at the end of each arm. The training trial was started following the pretraining. During the training trial, each mouse was given two trials daily for 5 days. The same two arms were baited each day and across sessions; thus mice learned to ignore the remaining six arms that never contained a reward. A trial consisted of placing the mouse in the maze where it remained until both of the two reinforcements had been received or until 5 min had elapsed, whichever occurred first. Entry into a never-baited arm was considered a reference memory error, and choices of arms were recorded. Memory retention trial was conducted 3 days after the last training trial in the reference memory test of RAM task.

Y-maze

Short-term spatial memory was assessed by recording spontaneous alternation behavior in a Y-maze, which does not involve any training, reward, or punishment. The ability to alternate requires mice to know which arms have already been visited. Therefore, alternation behavior can be regarded as a measure involving spatial working memory. Each mouse was placed in the center of the symmetrical Y-maze and was allowed to explore freely through the maze during an 8-min session. The sequence and total number of arms entered were recorded. Arm entry was considered to be complete when the hind paws of the mouse had been completely placed in the arm. An alternation was defined as entries into all three arms on consecutive occasions. The number of maximum alternation was therefore the total number of arm entries minus 2, and the percentage of alternation was calculated as (actual alternations / maximum alternations) × 100.

Tissue processing

After behavioral tasks, mice were deeply anesthetized with a mixture of ketamine and xylazine and then decapitated. Isolated brains were bisected longitudinally and hemispheres were separated. Hippocampal tissues were dissected from the left hemisphere and frozen on dry ice for biochemical studies. The right hemisphere was kept in 4% paraformaldehyde (PFA) in cold 0.1 M phosphate buffer (pH 7.4) for 3 days, followed by incubation in 30% sucrose solution for immunostaining. Mice injected with BrdU were deeply anesthetized with a mixture of ketamine and xylazine, and perfused transcardially with 4% PFA in cold 0.1 M phosphate buffer (pH 7.4) after 0.9% NaCl. The brains were postfixed overnight and then transferred into a 30% sucrose solution and kept there until they sank. For immunostaining, 40-μm coronal sections were cut from a dry ice-cooled block on a sliding microtome (Leica, Wetzlar, Germany). The sections were stored at −20°C in a cryoprotective buffer containing 28% ethylene glycol, 23% glycol, and 0.05 M phosphate buffer until processing for immunohistochemistry or immunofluorescence.

Immunohistochemistry and immunofluorescence confocal microscopy

Immunohistochemistry and immunofluorescent labeling were performed as described previously (*54*). The antibodies used were rat anti-BrdU (1:100, Accurate Chemical & Scientific Corporation, Westbury, NY), biotin-conjugated mouse anti-BrdU (1:100, Millipore, Temecula, CA), goat anti-DCX (1:200, Santa Cruz Biotechnology, Dallas, Texas), mouse anti-neuronal nuclei (NeuN, 1:500, Chemicon, Temecula, CA), rabbit anti-Fox3/NeuN (1:500, EnCor Biotechnology Inc., Gainesville, FL), rabbit anti-Glial Fibrillary Acidic Protein (GFAP, 1:500, Dako, Fort Collins, CO), rabbit anti-ionized calcium-binding adaptor molecule 1 (Iba1, 1:500, Wako, Osaka, Japan), goat anti-Iba1 (1:500, Abcam, Cambridge, MA), mouse anti-Aβ 3D6 antibodies (1:2,500, a gift from Lilly), and rabbit monoclonal anti-cleaved Caspase 3 (Casp3, 1:1,000, Cell signaling, Danvers, MA).

The immunohistochemical staining was made using the avidin-biotin complex (ABC) system and nickel-enhanced diaminobenzidine (DAB) incubation (Vectastain Elite, Vector labs, Burlingame, CA). Sections were mounted on gelatin-coated slides, air-dried, dehydrated, cleared, and coverslipped. The fluorescent secondary antibodies used for immunofluorescent labeling were donkey anti-goat immunoglobulin G (IgG) conjugated with Cy2 or Cy3; donkey anti-rat IgG conjugated with Cy2 or Cy3; donkey anti-mouse IgG conjugated with Cy2 or Cy3 or Cy5; donkey anti-rabbit IgG conjugated with Cy2 or Cy3 or Cy5 (all 1:250, Jackson ImmunoResearch, West Grove, PA). Fluorescent signals were detected using a LSM Pascal 5 Carl Zeiss confocal laser scanning microscope (Zeiss, Germany). Images were captured and recorded using a Zeiss LSM image browser.

For BrdU staining, DNA was denatured by incubating the sections for 2 hours in the 50%

formamide/2× SSC (0.3 M NaCl and 0.03 M sodium citrate) at 65°C. Sections were rinsed for 15 min in 2× SSC and incubated for 30 min in 2 N HCl at 37°C. Acid was neutralized by rinsing the sections for 10 min in 0.1 M boric acid (pH 8.5) followed by several washes in Tris-buffered saline (TBS, pH 7.5).

Quantification of immunoreactivity and mouse regrouping

For estimating total $BrdU^+$ cells, a series of systematically selected every sixth brain section was stained, and $BrdU^+$ cells in the subgranular cell layer (SGL) and granular cell layer (GCL) were counted by collecting images under 40× objective on a light microscope (TE360 Eclipse, Nikon, Japan) or a confocal microscope (LSM Pascal 5 Carl Zeiss, Germany). The sum of the $BrdU^+$ cell counts was multiplied by 6 to obtain an estimate of total numbers. For co-labeling analysis of differentiated $BrdU^+$ cell types with lineage-specific markers, the phenotypes of 30 $BrdU^+$ cells per animal were determined. Both the $BrdU^+$ cell count and the phenotype analysis were performed blindly.

For quantification of DCX^+ cells, a series of systematically selected every 6th or 12th brain section from each mouse was taken from similar regions spanning between bregma −1.34 mm and −3.28 mm using a Mouse Brain Atlas [Paxinos G, Franklin KBJ (2001) The mouse brain in stereotaxic coordinates, Academic Press, San Diego]. The numbers of DCX^+ cells were counted in the inner rim, defined as the border between the hilus and GCL, under 20× objective lens manually. The inner rim length of GCL was measured using a Zeiss LSM image browser (Carl Zeiss, Germany) or Imaris software (Bitplane, Concord, MA). Based on the results of DCX^+ cell numbers, mice were regrouped for additional behavioral data analysis and biochemical/immunostaining experiments. See tables S1 and S2, and fig. S18A for the number of animals in each experimental group and explanations of group arrangement.

For quantification of $Casp3^+$ cells, a series of systematically selected every sixth brain section was stained, and $Casp3^+$ cells in the SGL and GCL were counted by collecting images under 40× objective on a light microscope (TE360 Eclipse, Nikon, Japan) or a confocal microscope (LSM Pascal 5 Carl Zeiss, Germany). The sum of the $Casp3^+$ cell counts was multiplied by 6 to obtain an estimate of total numbers. Researchers who were unaware of the experimental group to which the samples belonged quantified DCX^+ or $Casp3^+$ cells.

Total granule cell counting

Granule cell numbers were determined using unbiased stereologic counting methods. Each sixth section containing the hippocampus was Nissl stained. Optical disector frames (15 × 15 μm) were set in as systematic-random fashion, accounting for 1% of the area of the GCL on each section. Cells were counted in 5-μm-thick stacks of optical disectors (1 μm in depth for each disector), according to stereologic principles. The total number of GCL neurons was obtained by using the formula developed as described previously (*55*). The granule cell counting was performed blindly.

Quantitative assessment of amyloid deposition

To examine the impact of AHN manipulation or exercise on amyloid deposition, a series of systematically selected (every 12th) brain sections from each mouse was probed with Aβ-specific 3D6 antibodies, detected by fluorescently labeled secondary antibodies. Images to quantify the amyloid burden were collected in a Z series of 40-μm depth with 4-μm intervals between images. The volume of amyloid burden was quantified using ImageJ (Voxel counter plugin), and the area of Aβ immunoreactivity was assessed after adequate thresholding and isolated noise despeckling.

Quantitative cytokine level measurement and ELISAs

Hippocampal tissues were homogenized in RIPA buffer (Sigma, St. Louis, MO) and centrifuged at 45,000*g* for 30 min at 4°C. Supernatants were used to measure 10 cytokines: Interleukin 1β (IL-1β), IL-2, IL-4, IL-5, IL-6, IL-10, IL-12p70, tumor necrosis factor α (TNFα), interferon-γ (IFN-γ), and keratinocyte-derived chemokine (KC/GRO). The measurement of cytokines was performed using the MesoScale Discovery (MSD, Rockville, MD) 96-well Mouse Pro-Inflammatory V-PLEX Assay as outlined in the manufacturer's protocol. Briefly, 25 μl each of sample and calibrator were added to the plate coated with an array of cytokine capture antibodies. The plate was then incubated for 2 hours with vigorous shaking at room temperature, followed by washing with wash buffer (provided in the kit). A volume of 25 μl of the detection antibody solution was added and incubated for 2 hours with vigorous shaking at room temperature. The plate was washed with wash buffer before adding 150 μl 2× MSD Read Buffer, and immediately read on a Meso QuickPlex SQ 120.

For TGF-β1, PSD95, SAP97, SYP, and FNDC5 analyses, supernatant samples were further diluted in cold PBS. For TGF-β1, diluted samples were acidified to pH < 3 with 1 N HCl and incubated for 1 hour at room temperature, followed by neutralization to pH ~8 with 1 N NaOH. Levels of TGF-β1 were measured using TGF-β1 Platinum ELISA kits (Affymetrix eBioscience, San Diego, CA) according to the manufacturer's protocols. PSD95, SAP97, SYP, and FNDC5 contents were measured using ELISA kits from MyBioSource (San Diego, CA) according to the manufacturer's protocols. BDNF protein levels were measured using BDNF ELISA kits (R&D Systems, Minneapolis, MN). TGF-β1 in the media of the 3D cultures were measured using the MSD 96-well TGF-β1 kit (Rockville, MD).

Immunoblot analysis

Fifteen to 75 μg of protein were resolved on 12% Bis-Tris or 4 to 12% gradient Bis/Tris gels (Life Technologies, Grand Island, NY), and the proteins were transferred to the nylon membranes (Bio-Rad, Hercules, CA). Immunoblot images were visualized by enhanced chemiluminescence (ECL). The images were captured by using BioMax film (Kodak, Rochester, NY) or VersaDoc imaging system (Bio-Rad, Hercules, CA) and quantitated by using QuantityOne software (Bio-Rad, Hercules, CA). Primary antibodies were used at the following dilutions: anti-PSD95 (1:500, NeuroMab, Davis, CA), anti-synapse-associated protein 97 (SAP97, 1:500, NeuroMab, Davis, CA), and anti-synaptophysin (SYP, 1:1,000, Millipore, Temecula, CA).

Golgi staining

Golgi staining was performed using the FD Rapid Golgi Stain Kit (FD Neurotechnologies, Inc., Columbia, MD), a simplified and reliable kit for Golgi impregnation, to label neurons in the outer layer of GCLs. Animals were deeply anesthetized with an isoflurane, and the brains were immediately removed and rinsed in MilliQ water. After the rinse, retrieved brains were immersed in a Golgi-Cox solution comprising potassium dichromate, mercuric chloride, and potassium chromate. This mixture of solutions was replaced once after 6 hours of initial immersion, with storage at room temperature in darkness for 2 weeks. After the immersion period in the Golgi-Cox solution, the embedded brains were transferred to a cryoprotectant solution (FD Rapid Golgi Stain Kit) and stored at room temperature for at least 72 hours in the dark before cutting. The brain slices were sectioned in the coronal or sagittal plane at approximately 250 μm thickness on a cryostat. Sliced tissues were transferred onto gelatin-coated slides and were air dried at room temperature in the dark. After drying, sections were rinsed with distilled water and were subsequently stained in a developing solution (FD Rapid Golgi Stain Kit) and dehydrated with 50, 75, 95, and 100% ethanol. Finally, the sections were defatted in xylene substitute and coverslipped with either Permount (Fisher Scientific, Fair Lawn, NJ).

Images were acquired from prepared slides using a Zeiss 510 microscope. Each neuron was scanned under high (100×, oil immersion) magnification by varying the depth of the Z plane to ensure that all parts of the cell (especially dendrites) were intact. Dendrites that tapered to a point were assumed to be complete and uncut. At least 20 neurons were selected. 3D neuronal reconstruction was performed by LSM510 (Zeiss). The total length and number of branches of dendrites and the spine density were measured by Imaris software (Bitplane, Concord, MA). Seven neurons per mouse were examined.

Treatment of 3D FAD ReN cell cultures with TGF-β1

ReNcell VM human NPCs (ReN cells, EMD Millipore, Billerica, MA) were transfected with internal ribosome entry site (IRES)–mediated polycistronic lentiviral vectors containing FAD genes encoding human APP with both Swedish (K670N/M671L) and London (V717I) mutations, or PS1ΔE9

mutation along with APP Swedish/London mutations, with GFP as a reporter for viral infection. Fluorescence-activated cell sorting (FACS) was then used to enrich the population of cells with the highest expression levels. ReN cells with GFP alone served as controls. The FACS-sorted ReN cells expressing high levels of FAD genes were differentiated and maintained in a 3D Matrigel culture system. Differentiation media containing either 1, 5, or 10 ng/ml of active human TGF-β1 (Abcam, Cambridge, MA) were added to 6-week differentiated cultures every 3 days, and the cells were maintained for an additional 2 weeks. The cell media were collected for Lactate dehydrogenase (LDH) assay, CellTiter-Glo luminescent cell viability assay and Aβ quantification. The cells were fixed with 4% PFA for Casp3 staining.

LDH assay

CytoTox-ONE assay was performed according to the manufacturer's guidelines (Promega, Madison, WI). Fifty microliters of cell culture medium was removed from the cells and substrate was added to the cell culture medium in 1:1 dilution and incubated for 30 min in a 37°C incubator. The plate was measured using a spectrophotometer (excitation: 560 nm, emission: 590 nm). Under the influence of the assay's substrate, resazurin is converted to the fluorescent form resorufin due to LDH, which is released into the medium by dead cells only. Therefore, increased values during the experiments were interpreted as increased cell death.

CellTiter-Glo luminescent cell viability assay

Cell viability was determined using the CellTiter-Glo luminescent cell viability kit from Promega Corporation (Madison, WI) according to the manufacturer's instructions. This method was based on the measurement of ATP production in the cells, which is proportional to the number of viable cells, detected by luciferin-luciferase reaction. One hundred microliters of CellTiter-Glo Reagent was added into the wells. Contents were mixed for 2 min on an orbital shaker to induce cell lysis. The plate was allowed to incubate at room temperature for 10 min to stabilize luminescent signal, and luminescence was recorded.

Casp3 staining using 3D cultures

The fixed cells were permeabilized and blocked by incubating with a blocking solution containing 50 mM Tris (pH 7.4), 0.1% Tween-20, 4% donkey serum, 1% BSA, 0.1% gelatin, and 0.3 M glycine at 4°C for 12 hours. After washing with TBS buffer containing 0.1% (v/v) Tween-20 (TBST), the 3D cultures were incubated with rabbit monoclonal anti-cleaved Casp3 (1:1,000, Cell signaling, Danvers, MA) antibodies in the blocking solution at 4°C overnight. After washing three times with TBST, the cells were incubated with donkey anti-rabbit IgG conjugated Cy3 for 2 hours at room temperature (1:250, Jackson ImmunoResearch, West Grove, PA). To avoid fluorescence quenching, a drop of anti-fade gold (Life Technologies, Grand Island, NY) was added on top of the fixed/stained thin-layer 3D cultures before imaging.

Dot-blot analysis

3D Matrigel samples were dissolved in Lysis Buffer 6 (R&D Systems, Minneapolis, MN) with protease and phosphatase inhibitors (Thermo Scientific, Rockford, IL). One microliter of each sample was spotted onto a LI-COR Biosciences Odyssey Nitrocellulose Membrane (LI-COR Biosciences, Lincoln, NE). The dot blot was then stained for Monoclonal Anti-β-Actin-Peroxidase antibody produced in mouse (Sigma Aldrich, St. Louis, MO) overnight at a dilution ratio of 1:50,000. Then, after sufficient washing the following day, the blot was coated with 2 ml of Pierce ECL Western Blotting Substrate (Thermo Scientific, Rockford, IL). Chemiluminescent detection imaging was then conducted using the Chemi exposure setting on a Li-Cor Odyssey FC machine (LI-COR Biosciences, Lincoln, NE) and then quantified using Image Studio software (LI-COR Biosciences, Lincoln, NE).

Statistical analysis

Data are expressed as mean values ± standard error of the mean (mean ± SEM). Error bars in the figures represent SEM. Differences between groups were analyzed using a *t*-test or an ANOVA that was followed by a post hoc comparison using Tukey or Fisher's least significant difference (LSD), where appropriate. For the RAM tests, a two-way repeated measures analysis of ANOVA was used for main effects (groups) with day as the repeated measure and errors as the dependent variable. In all cases, values of $P < 0.05$ were considered to be significant, and * or # indicated $P < 0.05$ and ** or ## indicated $P < 0.01$. Statistical analysis was performed using PRISM GraphPad statistical software.

REFERENCES AND NOTES

1. R. E. Tanzi, L. Bertram, Twenty years of the Alzheimer's disease amyloid hypothesis: A genetic perspective. *Cell* **120**, 545–555 (2005). doi: 10.1016/j.cell.2005.02.008; pmid: 15734686
2. P. S. Eriksson *et al*., Neurogenesis in the adult human hippocampus. *Nat. Med.* **4**, 1313–1317 (1998). doi: 10.1038/3305; pmid: 9809557
3. Y. Mu, F. H. Gage, Adult hippocampal neurogenesis and its role in Alzheimer's disease. *Mol. Neurodegener.* **6**, 85–93 (2011). doi: 10.1186/1750-1326-6-85; pmid: 22192775
4. L. Crews *et al*., Increased BMP6 levels in the brains of Alzheimer's disease patients and APP transgenic mice are accompanied by impaired neurogenesis. *J. Neurosci.* **30**, 12252–12262 (2010). doi: 10.1523/JNEUROSCI.1305-10.2010; pmid: 20844121
5. D. Gomez-Nicola *et al*., Temporal dynamics of hippocampal neurogenesis in chronic neurodegeneration. *Brain* **137**, 2312–2328 (2014). doi: 10.1093/brain/awu155; pmid: 24941947
6. K. Jin *et al*., Increased hippocampal neurogenesis in Alzheimer's disease. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 343–347 (2004). doi: 10.1073/pnas.2634794100; pmid: 14660786
7. B. Li *et al*., Failure of neuronal maturation in Alzheimer disease dentate gyrus. *J. Neuropathol. Exp. Neurol.* **67**, 78–84 (2008). doi: 10.1097/nen.0b013e318160c5db; pmid: 18091557
8. M. A. Lovell, H. Geiger, G. E. Van Zant, B. C. Lynn, W. R. Markesbery, Isolation of neural precursor cells from Alzheimer's disease and aged control postmortem brain. *Neurobiol. Aging* **27**, 909–917 (2006). doi: 10.1016/j.neurobiolaging.2005.05.004; pmid: 15979211
9. E. K. Perry *et al*., Neurogenic abnormalities in Alzheimer's disease differ between stages of neurogenesis and are partly related to cholinergic pathology. *Neurobiol. Dis.* **47**, 155–162 (2012). doi: 10.1016/j.nbd.2012.03.033; pmid: 22504537
10. H. Oakley *et al*., Intraneuronal beta-amyloid aggregates, neurodegeneration, and neuron loss in transgenic mice with five familial Alzheimer's disease mutations: Potential factors in amyloid plaque formation. *J. Neurosci.* **26**, 10129–10140 (2006). doi: 10.1523/JNEUROSCI.1202-06.2006; pmid: 17021169
11. O. Lazarov *et al*., Environmental enrichment reduces Abeta levels and amyloid deposition in transgenic mice. *Cell* **120**, 701–713 (2005). doi: 10.1016/j.cell.2005.01.015; pmid: 15766532
12. Z. Radak *et al*., Exercise plays a preventive role against Alzheimer's disease. *J. Alzheimers Dis.* **20**, 777–783 (2010). doi: 10.3233/JAD-2010-091531; pmid: 20182027
13. H. van Praag, G. Kempermann, F. H. Gage, Running increases cell proliferation and neurogenesis in the adult mouse dentate gyrus. *Nat. Neurosci.* **2**, 266–270 (1999). doi: 10.1038/6368; pmid: 10195220
14. A. A. Pieper *et al*., Discovery of a proneurogenic, neuroprotective chemical. *Cell* **142**, 39–51 (2010). doi: 10.1016/j.cell.2010.06.018; pmid: 20603013
15. D. C. Lie *et al*., Wnt signalling regulates adult hippocampal neurogenesis. *Nature* **437**, 1370–1375 (2005). doi: 10.1038/nature04108; pmid: 16251967
16. C. D. Clelland *et al*., A functional role for adult hippocampal neurogenesis in spatial pattern separation. *Science* **325**, 210–213 (2009). doi: 10.1126/science.1173215; pmid: 19590004
17. L. Deuker, C. F. Doeller, J. Fell, N. Axmacher, Human neuroimaging studies on the hippocampal CA3 region - integrating evidence for pattern separation and completion. *Front. Cell. Neurosci.* **8**, 64 (2014). doi: 10.3389/fncel.2014.00064; pmid: 24624058
18. J. Götz, L. M. Ittner, Animal models of Alzheimer's disease and frontotemporal dementia. *Nat. Rev. Neurosci.* **9**, 532–544 (2008). doi: 10.1038/nrn2420; pmid: 18568014
19. D. Meshi *et al*., Hippocampal neurogenesis is not required for behavioral effects of environmental enrichment. *Nat. Neurosci.* **9**, 729–731 (2006). doi: 10.1038/nn1696; pmid: 16648847
20. L. Santarelli *et al*., Requirement of hippocampal neurogenesis for the behavioral effects of antidepressants. *Science* **301**, 805–809 (2003). doi: 10.1126/science.1083328; pmid: 12907793
21. A. Garthe, J. Behr, G. Kempermann, Adult-generated hippocampal neurons allow the flexible use of spatially precise learning strategies. *PLOS ONE* **4**, e5464 (2009). doi: 10.1371/journal.pone.0005464; pmid: 19421325
22. C. E. Finch, N. J. Laping, T. E. Morgan, N. R. Nichols, G. M. Pasinetti, TGF-beta 1 is an organizer of responses to neurodegeneration. *J. Cell. Biochem.* **53**, 314–322 (1993). doi: 10.1002/jcb.240530408; pmid: 8300749
23. K. C. Flanders, R. F. Ren, C. F. Lippa, Transforming growth factor-betas in neurodegenerative disease. *Prog. Neurobiol.* **54**, 71–85 (1998). doi: 10.1016/S0301-0082(97)00066-X; pmid: 9460794
24. H. J. Klassen *et al*., Expression of cytokines by multipotent neural progenitor cells. *Cytokine* **22**, 101–106 (2003). doi: 10.1016/S1043-4666(03)00120-0; pmid: 12849709
25. Y. Zhu, B. Ahlemeyer, E. Bauerbach, J. Krieglstein, TGF-beta1 inhibits caspase-3 activation and neuronal apoptosis in rat hippocampal cultures. *Neurochem. Int.* **38**, 227–235 (2001). doi: 10.1016/S0197-0186(00)00084-X; pmid: 11099781
26. D. Battista, C. C. Ferrari, F. H. Gage, F. J. Pitossi, Neurogenic niche modulation by activated microglia: Transforming growth factor beta increases neurogenesis in the adult dentate gyrus. *Eur. J. Neurosci.* **23**, 83–93 (2006). doi: 10.1111/j.1460-9568.2005.04539.x; pmid: 16420418
27. M. S. Buckwalter *et al*., Chronically increased transforming growth factor-beta1 strongly inhibits hippocampal neurogenesis in aged mice. *Am. J. Pathol.* **169**, 154–164 (2006). doi: 10.2353/ajpath.2006.051272; pmid: 16816369
28. M. Kandasamy *et al*., TGF-beta signalling in the adult neurogenic niche promotes stem cell quiescence as well as generation of new neurons. *J. Cell. Mol. Med.* **18**, 1444–1459 (2014). doi: 10.1111/jcmm.12298; pmid: 24779367
29. F. P. Wachs *et al*., Transforming growth factor-beta1 is a negative modulator of adult neurogenesis. *J. Neuropathol. Exp. Neurol.* **65**, 358–370 (2006). doi: 10.1097/01.jnen.0000218444.53405.f0; pmid: 16691117

30. S. H. Choi *et al*., A three-dimensional human neural cell culture model of Alzheimer's disease. *Nature* **515**, 274–278 (2014). doi: 10.1038/nature13800; pmid: 25307057
31. J. Cho *et al*., Treadmill Running Reverses Cognitive Declines due to Alzheimer Disease. *Med. Sci. Sports Exerc.* **47**, 1814–1824 (2015). doi: 10.1249/MSS.0000000000000612; pmid: 25574797
32. C. W. Cotman, N. C. Berchtold, Exercise: A behavioral intervention to enhance brain health and plasticity. *Trends Neurosci.* **25**, 295–301 (2002). doi: 10.1016/S0166-2236(02)02143-4; pmid: 12086747
33. C. W. Cotman, N. C. Berchtold, L. A. Christie, Exercise builds brain health: Key roles of growth factor cascades and inflammation. *Trends Neurosci.* **30**, 464–472 (2007). doi: 10.1016/j.tins.2007.06.011; pmid: 17765329
34. M. Llorens-Martín, I. Torres-Alemán, J. L. Trejo, Growth factors as mediators of exercise actions on the brain. *Neuromolecular Med.* **10**, 99–107 (2008). doi: 10.1007/s12017-008-8026-1; pmid: 18286390
35. S. S. Vaynman, Z. Ying, D. Yin, F. Gomez-Pinilla, Exercise differentially regulates synaptic proteins associated to the function of BDNF. *Brain Res.* **1070**, 124–130 (2006). doi: 10.1016/j.brainres.2005.11.062; pmid: 16413508
36. C. D. Wrann *et al*., Exercise induces hippocampal BDNF through a PGC-1α/FNDC5 pathway. *Cell Metab.* **18**, 649–659 (2013). doi: 10.1016/j.cmet.2013.09.008; pmid: 24120943
37. J. E. Hamos, L. J. DeGennaro, D. A. Drachman, Synaptic loss in Alzheimer's disease and other dementias. *Neurology* **39**, 355–361 (1989). doi: 10.1212/WNL.39.3.355; pmid: 2927643
38. S. Peng, J. Wuu, E. J. Mufson, M. Fahnestock, Precursor form of brain-derived neurotrophic factor and mature brain-derived neurotrophic factor are decreased in the pre-clinical stages of Alzheimer's disease. *J. Neurochem.* **93**, 1412–1421 (2005). doi: 10.1111/j.1471-4159.2005.03135.x; pmid: 15935057
39. D. T. Proctor, E. J. Coulson, P. R. Dodd, Reduction in post-synaptic scaffolding PSD-95 and SAP-102 protein levels in the Alzheimer inferior temporal cortex is correlated with disease pathology. *J. Alzheimers Dis.* **21**, 795–811 (2010). doi: 10.3233/JAD-2010-100090; pmid: 20634587
40. P. Rasmussen *et al*., In humans IL-6 is released from the brain during and after exercise and paralleled by enhanced IL-6 mRNA expression in the hippocampus of mice. *Acta Physiol. (Oxf.)* **201**, 475–482 (2011). doi: 10.1111/j.1748-1716.2010.02223.x; pmid: 21083649
41. P. C. Baier, U. May, J. Scheller, S. Rose-John, T. Schiffelholz, Impaired hippocampus-dependent and -independent learning in IL-6 deficient mice. *Behav. Brain Res.* **200**, 192–196 (2009). doi: 10.1016/j.bbr.2009.01.013; pmid: 19378383
42. K. K. Bowen, R. J. Dempsey, R. Vemuganti, Adult interleukin-6 knockout mice show compromised neurogenesis. *Neuroreport* **22**, 126–130 (2011). doi: 10.1097/WNR.0b013e3283430a44; pmid: 21266900
43. C. A. McPherson, M. Aoyama, G. J. Harry, Interleukin (IL)-1 and IL-6 regulation of neural progenitor cell proliferation with hippocampal injury: Differential regulatory pathways in the subgranular zone (SGZ) of the adolescent and mature mouse brain. *Brain Behav. Immun.* **25**, 850–862 (2011). doi: 10.1016/j.bbi.2010.09.003; pmid: 20833246
44. L. Vallières, I. L. Campbell, F. H. Gage, P. E. Sawchenko, Reduced hippocampal neurogenesis in adult transgenic mice with chronic astrocytic production of interleukin-6. *J. Neurosci.* **22**, 486–492 (2002). doi: 10.1523/JNEUROSCI.22-02-00486.2002; pmid: 11784794
45. R. Yirmiya, I. Goshen, Immune modulation of learning, memory, neural plasticity and neurogenesis. *Brain Behav. Immun.* **25**, 181–213 (2011). doi: 10.1016/j.bbi.2010.10.015; pmid: 20970492
46. C. Zhao, W. Deng, F. H. Gage, Mechanisms and functional implications of adult neurogenesis. *Cell* **132**, 645–660 (2008). doi: 10.1016/j.cell.2008.01.033; pmid: 18295581
47. D. Guerrieri, H. van Praag, Exercise-mimetic AICAR transiently benefits brain function. *Oncotarget* **6**, 18293–18313 (2015). doi: 10.18632/oncotarget.4715; pmid: 26286955
48. T. Kobilo *et al*., AMPK agonist AICAR improves cognition and motor coordination in young and aged mice. *Learn. Mem.* **21**, 119–126 (2014). doi: 10.1101/lm.033332.113; pmid: 24443745
49. X. P. Hong *et al*., Essential role of tau phosphorylation in adult hippocampal neurogenesis. *Hippocampus* **20**, 1339–1349 (2010). doi: 10.1002/hipo.20712; pmid: 19816983
50. H. Hörder *et al*., Midlife cardiovascular fitness and dementia: A 44-year longitudinal population study in women. *Neurology* **90**, e1298–e1305 (2018). doi: 10.1212/WNL.0000000000005290; pmid: 29540588
51. S. E. Lamb *et al*., Dementia And Physical Activity (DAPA) trial of moderate to high intensity exercise training for people with dementia: Randomised controlled trial. *BMJ* **361**, k1675 (2018). doi: 10.1136/bmj.k1675; pmid: 29769247
52. A. Santos-Lozano *et al*., Physical Activity and Alzheimer Disease: A Protective Association. *Mayo Clin. Proc.* **91**, 999–1020 (2016). doi: 10.1016/j.mayocp.2016.04.024; pmid: 27492909
53. D. Briley *et al*., Preserved neurogenesis in non-demented individuals with AD neuropathology. *Sci. Rep.* **6**, 27812 (2016). doi: 10.1038/srep27812; pmid: 27298190
54. S. H. Choi *et al*., Non-cell-autonomous effects of presenilin 1 variants on enrichment-mediated hippocampal progenitor cell proliferation and differentiation. *Neuron* **59**, 568–580 (2008). doi: 10.1016/j.neuron.2008.07.033; pmid: 18760694
55. M. J. West, H. J. Gundersen, Unbiased stereological estimation of the number of neurons in the human hippocampus. *J. Comp. Neurol.* **296**, 1–22 (1990). doi: 10.1002/cne.902960102; pmid: 2358525

ACKNOWLEDGMENTS

Funding: Supported by the Cure Alzheimer's Fund (R.E.T. and S.H.C.); the JPB Foundation (R.E.T., B.M.S., and F.H.G.); NIA and NIMH grant 5R01MH060009 (R.E.T.); The Henry and Allison McCance Center for Brain Health at MGH (R.E.T.); NIH/NIA grant 1RF1AG048080-01 (R.E.T. and D.Y.K.); NIH/NIA grant 2R01AG014713 (D.Y.K.); and the Mather's Foundation and the Leona M. and Harry B. Helmsley Charitable Trust (F.H.G.).
Author contributions: R.E.T. and S.H.C. initiated the project; R.E.T. supervised research; R.E.T., S.H.C., F.H.G., H.vP., and B.M.S. designed experiments; S.H.C., E.B., Z.K.C., C.A., and M.K.O. performed most of the research; S.W.L., G.D.C., and F.H.G. provided LV-Wnt3, LV-dnWnt, and LV-GFP; S.H.C. performed stereotaxic injections; B.P. and J.W.C. performed irradiation; E.B., J.A., and D.Y.K. performed 3D culture experiments; E.K. and A.R. performed Golgi staining, dot-blot, and ELISAs; M.K.O. and A.L. genotyped and maintained mice; C.Z. and S.J.M. performed MSD assays; and S.H.C., E.B., Z.K.C., and R.E.T. wrote the manuscript with editing and input from J.W.C., H.v.P., and F.H.G.
Competing interests: We declare no conflict of interest.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6406/eaan8821/suppl/DC1
Figs. S1 to S23
Tables S1 to S10

1 June 2017; resubmitted 4 June 2018
Accepted 17 July 2018
10.1126/science.aan8821

RESEARCH ARTICLE SUMMARY

STRUCTURAL BIOLOGY

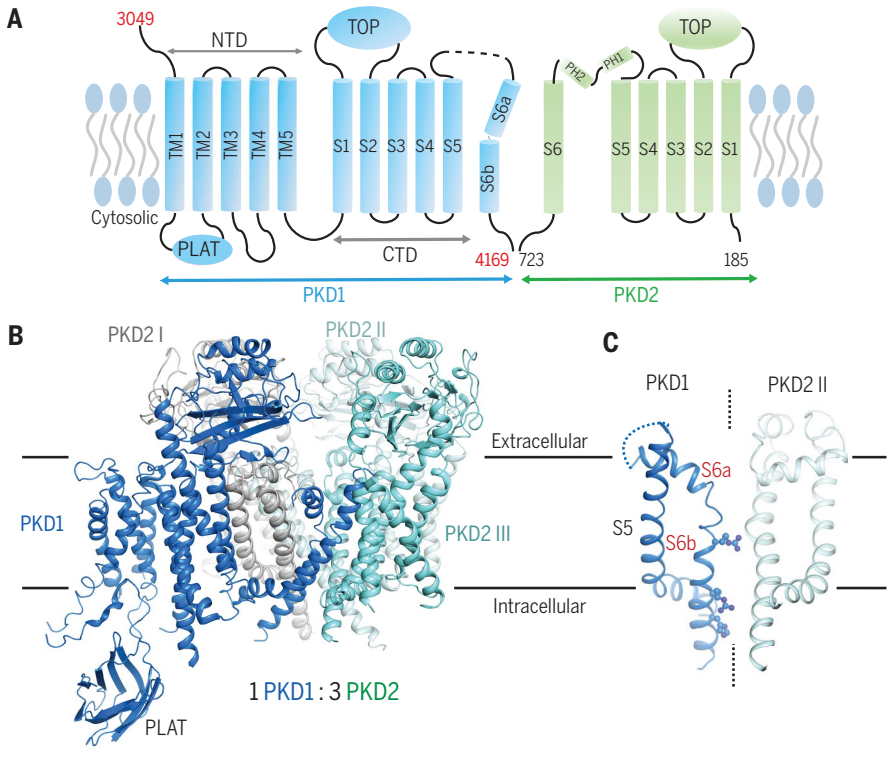
Structure of the human PKD1-PKD2 complex

Qiang Su*, Feizhuo Hu*, Xiaofei Ge, Jianlin Lei, Shengqiang Yu, Tingliang Wang, Qiang Zhou, Changlin Mei, Yigong Shi†

INTRODUCTION: Mutations in two genes, *PKD1* and *PKD2*, are responsible for about 85 and 10% of all cases of autosomal dominant polycystic kidney disease (ADPKD), one of the most common monogenetic disorders. However, the physiological and pathophysiological functions of the gene products polycystin-1 and polycystin-2 (PC1 and PC2, also known as PKD1 and PKD2) are not well understood.

PKD1, which comprises 4303 residues, may serve as a receptor that senses chemical and mechanical force stimuli, whereas PKD2, whose homotetrameric structure conforms to a typical group II transient receptor potential (TRP) channel, is hypothesized to be an endoplasmic reticulum Ca^{2+}-release channel and regulate intracellular Ca^{2+} concentrations. The two proteins were predicted to coexist as a hetero-oligomer on primary cilia in the renal epithelium, although the molecular basis for the formation of this complex remains elusive.

RATIONALE: To investigate the assembly of PKD1 and PKD2, we sought to resolve the structure of the PKD1-PKD2 complex. After extensive screening for optimal constructs and expression systems, a homogeneous complex was obtained through coexpression of FLAG-tagged PKD1 and Twin-Strep–tagged PKD2 (hereafter referred to as PKD1 and PKD2 for simplicity). Approximately 100 μg of the complex was obtained through affinity purification and size exclusion chromatography from 40 to 50 liters of suspension human embryonic kidney (HEK) 293F cells. The structure of the complex was determined to 3.6-Å resolution with single-particle cryo–electron microscopy (cryo-EM).



Cryo-EM structure of the truncated human PKD1-PKD2 complex at 3.6-Å resolution. (**A**) Topological illustration of PKD1 and PKD2. NTD, N-terminal domain; TOP, also known as the polycystin domain; CTD, C-terminal domain (which includes S1 to S6 and the TOP domain). (**B**) The 1:3 organization of the PKD1-PKD2 complex. PKD2 I, II, and III are the three PKD2 subunits. (**C**) Unconventional conformation of the S6 segment in PKD1. The sequences between the S5 and S6 segments are flexible and disordered in PKD1. The extracellular segment of the bent S6 resembles PH1.

RESULTS: PKD1 and PKD2 exhibit a 1:3 ratio in the structure. PKD1 consists of a voltage-gated ion channel (VGIC) fold that interacts with PKD2 to complete a domain-swapped TRP architecture. Several features, however, distinguish PKD1 from a canonical TRP channel. The S6 segment of PKD1 is broken in the middle, with the extracellular half, S6a, resembling pore helix 1 (PH1) in a typical VGIC. The sequence between S5 and S6a is highly flexible and disordered in the EM map. Three positively charged residues—Arg^{4100}, Arg^{4107}, and His^{4111}—protrude into the putative ion-conducting path, likely impeding permeability of the Ca^{2+} ion. Therefore, the current structure may represent a potentially nonconductive state.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aat9819

A discretely folded domain, which contains five transmembrane helices (TMs) and a cytosolic PLAT (polycystin-1, lipoxygenase, and alpha toxin) domain, precedes the VGIC fold in PKD1. The extracellular TOP domain of PKD1, which is frequently targeted for mutations in ADPKD, deviates from the expected symmetric position by 15°, leaving a gap in the extracellular TOP ring. Compared to the homotypic interactions among PKD2 subunits, the weakened interface between PKD1 and PKD2 provides a clue to the 1:3 stoichiometry in the heterotetramer. A higher ratio of PKD1 in the complex may weaken the association of the TOP domains.

CONCLUSION: The structure of the truncated PKD1-PKD2 complex reveals the molecular mechanism for the assembly of a hetero-oligomeric complex and provides a physical basis for mapping and understanding a large number of disease mutations. Elucidation of the functional mechanism of PKD1 and PKD2 as well as the disease mechanism of the hundreds of ADPKD mutations await further investigations. Our structure serves as a framework for future biophysical, biochemical, cellular, and computational analysis of PKD1-PKD2 and ADPKD. ■

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†Corresponding author. Email: shi-lab@tsinghua.edu.cn (Y.S.)
Cite this article as Q. Su *et al.*, *Science* **361**, eaat9819 (2018).
DOI: 10.1126/science.aat9819

RESEARCH ARTICLE

STRUCTURAL BIOLOGY

Structure of the human PKD1-PKD2 complex

Qiang Su[1]*, Feizhuo Hu[1]*, Xiaofei Ge[1], Jianlin Lei[2], Shengqiang Yu[3], Tingliang Wang[1,4], Qiang Zhou[1], Changlin Mei[3], Yigong Shi[1,4]†

Mutations in two genes, *PKD1* and *PKD2*, account for most cases of autosomal dominant polycystic kidney disease, one of the most common monogenetic disorders. Here we report the 3.6-angstrom cryo–electron microscopy structure of truncated human PKD1-PKD2 complex assembled in a 1:3 ratio. PKD1 contains a voltage-gated ion channel (VGIC) fold that interacts with PKD2 to form the domain-swapped, yet noncanonical, transient receptor potential (TRP) channel architecture. The S6 helix in PKD1 is broken in the middle, with the extracellular half, S6a, resembling pore helix 1 in a typical TRP channel. Three positively charged, cavity-facing residues on S6b may block cation permeation. In addition to the VGIC, a five–transmembrane helix domain and a cytosolic PLAT domain were resolved in PKD1. The PKD1-PKD2 complex structure establishes a framework for dissecting the function and disease mechanisms of the PKD proteins.

Autosomal dominant polycystic kidney disease (ADPKD) is a potentially lethal monogenetic disorder that manifests in renal disease and, in some cases, abnormalities in the liver, pancreas, brain, or arterial blood vessels (*1*). ADPKD is typically diagnosed in adults with an incidence of 1:400 to 1:1000 and affects about 6 million people worldwide (*2*). The hallmarks of the disease are bilateral, fluid-filled, enlarged renal cysts, which increase in number with age (*3*). Mutations in two genes, *PKD1* and *PKD2*, whose products are polycystin-1 and polycystin-2 (PC1 and PC2, also known as PKD1 and PKD2), account for about 85 and 10% of all ADPKD cases, respectively (*2*, *4*–*8*). Despite extensive effort, the physiological and pathophysiological mechanisms of PKD1 and PKD2 are not well understood.

PKD1 may function as a receptor that senses both chemical and mechanical force stimuli and regulates cytosolic cyclic adenosine monophosphate (cAMP) concentrations and downstream signaling (*6*, *9*–*12*). PKD2 is hypothesized to be an endoplasmic reticulum Ca^{2+}-release channel and regulate intracellular Ca^{2+} concentrations (*13*). In addition, PKD2 (colocalizes) with PKD1 on the shaft and basal body of primary cilia in the renal epithelium (*14*–*16*) and may contribute to fluid-flow sensation (*11*). Coexpression of human PKD1 and PKD2 is reported to produce distinct cation currents in Chinese hamster ovary cells (*17*). However, the channel activity of the ciliary PKD1-PKD2 complex is controversial. Delling *et al.* recently reported a complete lack of mechanically induced calcium influxes at physiological or supraphysiological levels of fluid flow against primary cilia (*18*).

Equally controversial is the molecular basis for the hetero-oligomerization of the PKD1-PKD2 complex. PKD1 and PKD2 were suggested to interact through their C-terminal coiled-coil domains (*19*–*22*). Other experiments, in which the complex was preserved in the absence of the coiled-coil domains, implied that complex formation may require the N-terminal loops (*23*–*25*). However, the cryo–electron microscopy (cryo-EM) structures of homotetrameric PKD2 and PKD2-like 1 protein (PKD2L1) revealed oligomerization in the absence of the coiled-coil domains or N-terminal loops (*4*, *26*). Furthermore, truncation of both the N- and C-terminal soluble domains of PKD2L1 did not alter the function of the PKD1L3-PKD2L1 complex, suggesting that these two elements are dispensable for hetero-oligomerization (*26*).

To elucidate the assembly of PKD1 and PKD2, we sought to resolve the structure of the PKD1-PKD2 complex. Several structures of PKD2 and related proteins in distinct states have been reported (*4*, *26*–*28*). In addition to the typical transient receptor potential (TRP) or voltage-gated ion channel (VGIC) transmembrane fold, each protomer contains an extracellular domain between the S1 and S2 segments that constitutes the TOP domain (also known as the polycystin domain), a feature shared by group II TRP channels (*29*). In contrast to the structural advances for PKD2, the only structural information on PKD1 is a nuclear magnetic resonance structure of a 78-residue PKD domain (residues 275 to 353) (*30*). The 4303-residue human PKD1 comprises an N-terminal extracellular region, 11 transmembrane helices (TMs), and a C-terminal coiled-coil domain (*6*). The extracellular segments form multiple domains involved in cell-cell or cell-matrix interactions (*7*). The transmembrane region can be divided into two entities: the N-terminal transmembrane domain (NTMD) containing five TMs and the C-terminal transmembrane domain (CTMD) that conforms to the VGIC fold (*6*). Within the NTMD, TM1 and TM2 are separated by a so-called PLAT (polycystin-1, lipoxygenase, and alpha toxin) domain (Fig. 1A). The highly conserved PLAT domain may participate in lipid binding and trafficking (*31*, *32*). Here we report the near-atomic-resolution cryo-EM structure of the complex between truncated human PKD1 and PKD2 in a closed conformation.

Results

Purification and structural determination of the PKD1-PKD2 complex

The bottleneck for structural elucidation of the PKD1-PKD2 complex was the expression and purification of homogeneous protein samples. To enhance biochemical stability, we screened numerous combinations of various constructs for both PKD1 and PKD2. Eventually, the complex obtained through coexpression of PKD1 residues 3049 to 4169 ($PKD1^{3049–4169}$) and $PKD2^{185–723}$ (hereafter referred to as PKD1 and PKD2 for simplicity), in which the putative flexible regions at the N and C termini of both proteins were removed (*27*), exhibited optimal solution behavior. Importantly, similar to full-length proteins, $PKD1^{3049–4169}$ is targeted to the cell surface only when coexpressed with $PKD2^{185–723}$ (fig. S1) (*33*).

The final yield of this complex was ~100 μg from 40 to 50 liters of suspension human embryonic kidney (HEK) 293F cell culture. To purify this low-yield protein complex, triple FLAG tag and Twin-Strep-tag were attached to the N-termini of PKD1 and PKD2, respectively (fig. S2A). After tandem affinity purification, PKD1 and PKD2 were monitored by Coomassie blue staining of SDS–polyacrylamide gel electrophoresis (PAGE) and verified by Western blotting and mass spectrometric (MS) analysis (fig. S2, B and C). After the last step of size exclusion chromatography purification, during which PKD1 and PKD2 comigrated, the peak fractions were pooled and concentrated for cryo-EM analysis (Fig. 1, B to D).

Details of sample preparation, data acquisition, and structural refinement can be found in the materials and methods. A total of 27,296 selected particles yielded a three-dimensional (3D) EM reconstruction at an overall resolution of 3.6 Å, according to the gold-standard Fourier shell correlation 0.143 criterion (*34*) (figs. S3 and S4). During local search 3D classification, the core region in the PKD1 subunit and two PKD2 subunits appeared more stable than the rest of the complex in some classes (fig. S4). We thereby combined these classes for further classification. Eventually 116,485 particles were selected to give rise to a map in which the resolution of

[1]Beijing Advanced Innovation Center for Structural Biology, Tsinghua-Peking Joint Center for Life Sciences, School of Life Sciences, School of Medicine, Tsinghua University, Beijing 100084, China. [2]Technology Center for Protein Sciences, Ministry of Education Key Laboratory of Protein Sciences, School of Life Sciences, Tsinghua University, Beijing 100084, China. [3]Department of Nephrology, Changzheng Hospital, Second Military Medical University, Shanghai 200433, China. [4]Institute of Biology, Westlake Institute for Advanced Study, Westlake University, 18 Shilongshan Road, Xihu District, Hangzhou 310064, Zhejiang Province, China.
*These authors contributed equally to this work.
†Corresponding author. Email: shi-lab@tsinghua.edu.cn (Y.S.)

this core region was improved to 3.2 Å (fig. S4 and table S1).

The "1 + 3" assembly of PKD1 and PKD2

The overall structure of the complex has dimensions of approximately 130 Å by 110 Å by 130 Å (Fig. 2A and fig. S5A). The transmembrane region of the complex comprises 29 TMs made up of two groups, the classical 24 TMs of an intact VGIC and 5 additional TMs. This organization is equivalent to a PKD2 homotetramer plus an NTMD of PKD1, confirming the presence of one PKD1 and three PKD2 (fig. S5B). The 1:3 stoichiometric ratio of PKD1 to PKD2 is consistent with single-molecule characterization of the full-length complex (*21*, *22*). Quantitative MS analysis also confirmed that the stoichiometry of the truncated PKD1-PKD2 complex is largely consistent with that of full-length proteins (*35*–*37*) (table S2).

To facilitate illustration, the 11 TMs in PKD1 will be described as $TM1^{PKD1}$ to $TM5^{PKD1}$ and $S1^{PKD1}$ to $S6^{PKD1}$. A β-barrel PLAT domain between $TM1^{PKD1}$ and $TM2^{PKD1}$ was resolved on the cytosolic side below the NTMD. The NTMD and PLAT are together described as the N-terminal domain (NTD) (Figs. 1A and 2B). Similar to PKD2, $S1^{PKD1}$ to $S4^{PKD1}$ constitute the voltage sensor–like domain (VSD^{PKD1}), and the sequence between S1 and S2 forms the extracellular TOP^{PKD1} domain (Fig. 2B) (*4*, *27*). The three PKD2 subunits are referred to as I, II, and III following a clockwise order in the extracellular view (Fig. 2A and fig. S5, A and B).

Most segments in the PKD1-PKD2 complex were well resolved in the map (fig. S5, C to E), enabling assignment of ~800 side chains for the VGIC region on the basis of the structure of PKD2 and a PKD2-derived PKD1 homologous model. An initial model for the NTD of PKD1, predicted by the I-TASSER (iterative threading assembly refinement) server (*38*), fit well into the corresponding EM map. The boundaries of the TMs in PKD1 are largely consistent with those derived from hydropathic predictions (table S3).

A distinctive S6 segment in PKD1

The three PKD2 protomers in the complex remain nearly identical to those in the PKD2 homotetramer (Fig. 3A and fig. S6, A and B). The fourfold symmetry of the pore domain (PD) in the PKD1-PKD2 complex is broken, owing to the distinct conformation of the $S6^{PKD1}$ segment (Fig. 3A and fig. S6B). The $S6^{PKD1}$ segment is bent in the middle, a feature that has not been observed in any other structures of VGIC-fold channels (*39*–*41*). The two halves, designated as S6a and S6b, form an axial angle of ~120° in the middle of the membrane (Fig. 3A, middle, and fig. S5D). Intriguingly, the conformation and position of the S6a segment are reminiscent of those of pore helix 1 (PH1) in a canonical TRP channel (Fig. 3A, right, and fig. S6B) (*29*), whereas the sequences corresponding to the putative selectivity filter (SF) and the supporting pore helices (PH1 and PH2) are invisible in PKD1 (Fig. 3A, middle, and fig. S6, B and C). Although the flexibility of this segment prevents detailed analysis, the conformation of S6a suggests that PKD1 may lack PH1, and also possibly lack a SF, when it is complexed with PKD2.

Another feature distinguishing $S6^{PKD1}$ from a typical VGIC or TRP channel is that it contains three cavity-facing, positively charged residues—Arg^{4100}, Arg^{4107}, and His^{4111}. These basic residues would be expected to disfavor cation penetration (Fig. 3B and fig. S6D), suggesting a potentially nonconductive state of the present structure. Arg^{4100} is stabilized by the neighboring Phe^{669} and Asn^{674} on S6 of PKD2 I ($S6_I$), and Arg^{4107} and His^{4111} interact with the polar residues Asn^{681} and Asp^{682} on $S6_I$ (Fig. 3B). The electrostatic interactions between $S6b^{PKD1}$ and $S6_I$ may provide the molecular basis for the 15° deviation of $S6^{PKD1}$ from the expected position on the basis of fourfold symmetry (Fig. 3C). Residue Arg^{4100} in PKD1 is highly conserved among different species, suggesting a critical role in the function of PKD1 (fig. S6C).

The structure of NTD^{PKD1}

A Dali (*42*) search suggests that the PKD1-NTMD represents a previously uncharacterized fold, in which the five TMs form a helical bundle and TM1 and TM2 are separated in the primary sequence by the PLAT domain (Fig. 4A). TM2 is preceded by a membrane reentrant loop (hereafter referred to as the pre-TM2 loop), which inserts into the cavity of the five-TM helical bundle, with a highly conserved Trp^{3263} interacting with

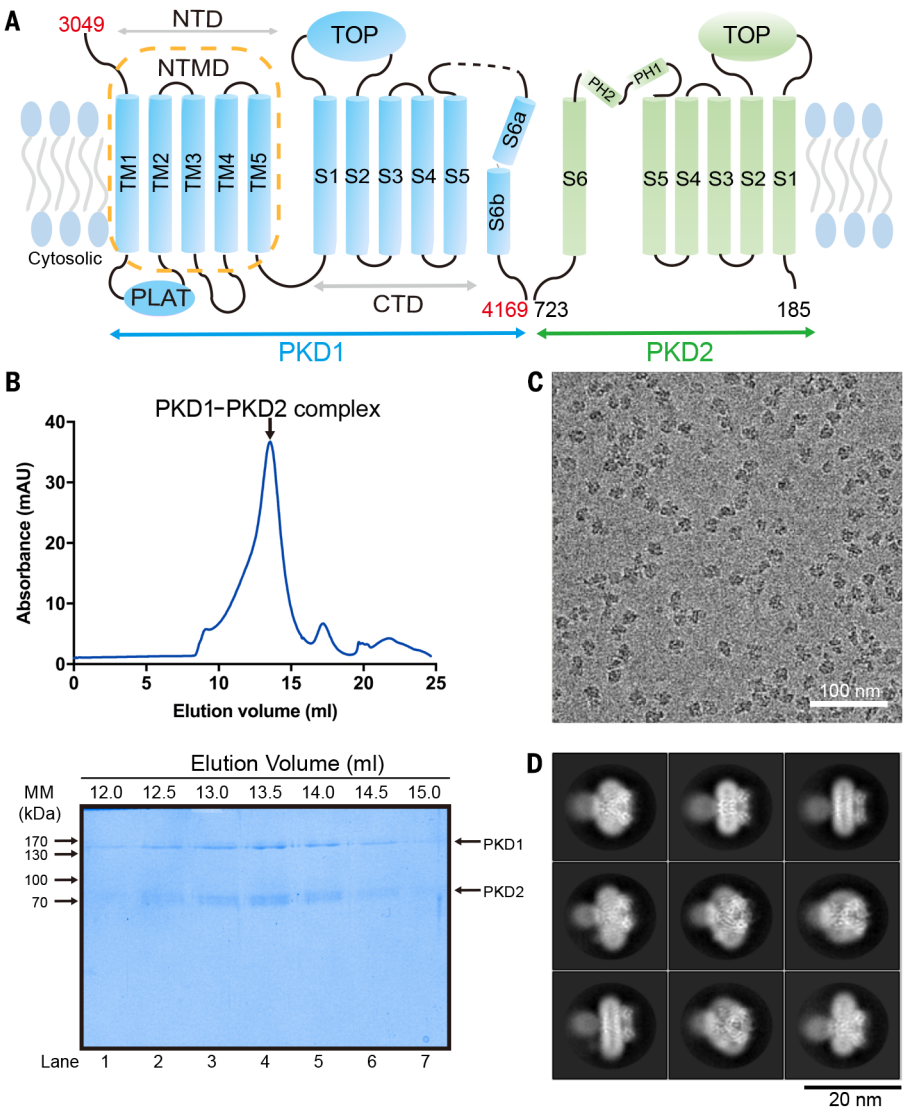


Fig. 1. Expression and purification of the human PKD1-PKD2 complex. (**A**) Topological illustration of PKD1 and PKD2. (**B**) Purification of the PKD1-PKD2 complex. Shown is a representative size exclusion chromatogram of the PKD1-PKD2 complex. The indicated peak fractions were resolved on SDS-PAGE and visualized by Coomassie blue staining. The smeary bands corresponding to PKD2 resulted from heterogeneous glycosylation. mAU, milli–arbitrary units. (**C**) Representative cryo-EM micrograph of the complex. MM, molecular mass. (**D**) 2D class averages of the complex.

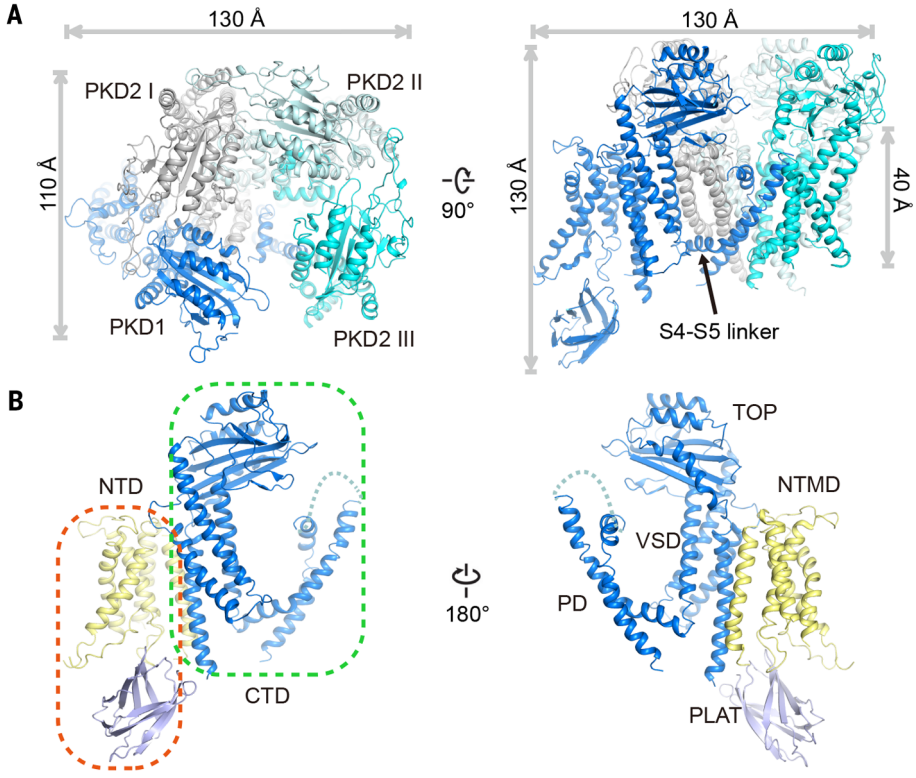


Fig. 2. The "1 + 3" organization of the PKD1 and PKD2 complex. (**A**) Overall structure of the complex. PKD1 is colored blue and the three PKD2 subunits are colored silver, pale cyan, and cyan. Two perpendicular views are shown. All structure figures were prepared in PyMol (*70*). (**B**) Structure of $PKD1^{3049–4169}$. Two opposing side views are shown, and the protomer is domain colored.

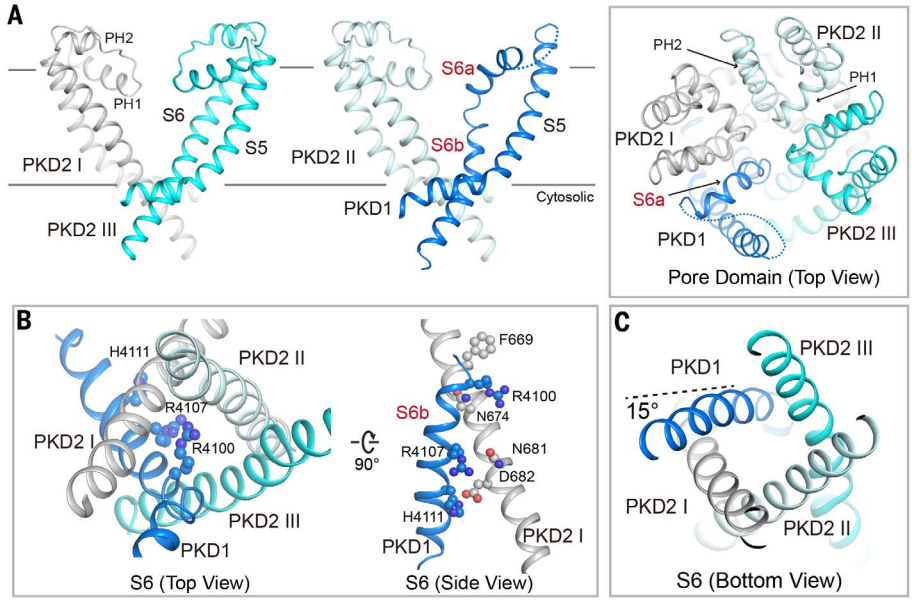


Fig. 3. Conformation of the $S6^{PKD1}$ segment. (**A**) PKD1 disrupts the fourfold symmetry of an otherwise typical VGIC fold. PKD1-S6 exhibits a distinct conformation from all VGIC channels of known structures. Whereas the sequences corresponding to the selectivity filter and the supporting helices (PH1 and PH2) are invisible in PKD1, the extracellular segment of the bent S6 resembles PH1. (**B**) Three positively charged residues on $S6b^{PKD1}$ may block cation permeation. Right: The conformation of $S6b^{PKD1}$ is stabilized by residues on PKD2-$S6_I$. The discussed residues are shown as spheres. (**C**) When viewed from the cytosolic side, $S6^{PKD1}$ displays a 15° deviation from the expected position for a fourfold symmetry.

Phe^{3596} and Phe^{3600} on TM5 (Fig. 4B and fig. S7A). This hydrophobic core may stabilize the overall conformation of the pre-TM2 loop. Missense mutation Trp^{3263}→Arg (W3263R) has been identified as a highly likely pathogenic mutation of ADPKD (*43*), perhaps because it is important for the structural integrity of NTMD. The interface between NTD and VSD^{PKD1} is constituted by hydrophobic residues on TM1 and S1 of PKD1 (Fig. 4C).

The four VSDs in the complex exhibit a pseudo fourfold symmetry. The conformations of the three VSDs from PKD2 are nearly identical (Fig. 4C and fig. S7B), whereas the overall structure of VSD^{PKD1} resembles that of the VSD^{PKD2} with minor variations. Compared to the corresponding segments in the VSDs of PKD2, the $S2^{PKD1}$ and $S3^{PKD1}$ segments move further away from the PD, whereas $S1^{PKD1}$ and $S4^{PKD1}$ shift toward the pore (fig. S7C). From the cytosolic view, the four helices in VSD^{PKD1} undergo a counterclockwise irislike rotation compared to those in VSD^{PKD2}, resulting in a larger intracellular mouth of the VSD^{PKD1} (fig. S7C). Whereas the S3 and S4 segments are connected by a short helix in PKD2, the linker helix becomes an extended fragment of S3 in PKD1, similar to that in the open-state structure of PKD2L1 (fig. S7C) (*26*).

Interactions between PKD1 and PKD2

The TOP domain in each protomer participates in the oligomerization of the homotetrameric PKD2 and PKD2L1 channels (*4*, *26*, *27*). By contrast, TOP^{PKD1} deviates by ~15° from the expected symmetric position relative to the three TOP^{PKD2} domains (Fig. 5A). When TOP^{PKD1} is superimposed on any TOP^{PKD2}, several structural distinctions are observed (Fig. 5B).

The most evident variation comes from the lack of the luminal loop from PKD1, which is missing in both the primary sequence and the 3D structure (fig. S8A) (*44*). In addition, the three-leaf clover (TLC) (*27*), also known as the finger 1 (*4*) motif, is invisible in the EM map of PKD1, likely owing to the flexibility of this region (Fig. 5B). The TLC is a critical element mediating the interactions of adjacent subunits in PKD2. The invisibility of TLC^{PKD1} may indicate the lack of stable interaction between TOP^{PKD1} and TOP_{III}, consistent with the 15° deviation of TOP^{PKD1}. On the other end, the TLC of TOP_I retains interaction with TOP^{PKD1}, an interface that is essential for complex assembly and trafficking (*45*) (fig. S8B).

The lack of a TLC^{PKD1}-mediated interface between PKD1 and PKD2 may provide a tentative clue to the 1:3 assembly of the heterotetramer. A higher ratio of PKD1 in the complex would lead to decreased association of the TOP domains, disfavoring higher occupancy of PKD1. Supporting the structural analysis, TOP^{PKD1} did not support homo-oligomerization of PKD1 in a co-immunoprecipitation assay (*45*).

The assembly of the PKD1-PKD2 complex involves two additional interfaces below the TOP domains. One is mediated by the electrostatic interaction between Arg^{3700} on TOP^{PKD1} and

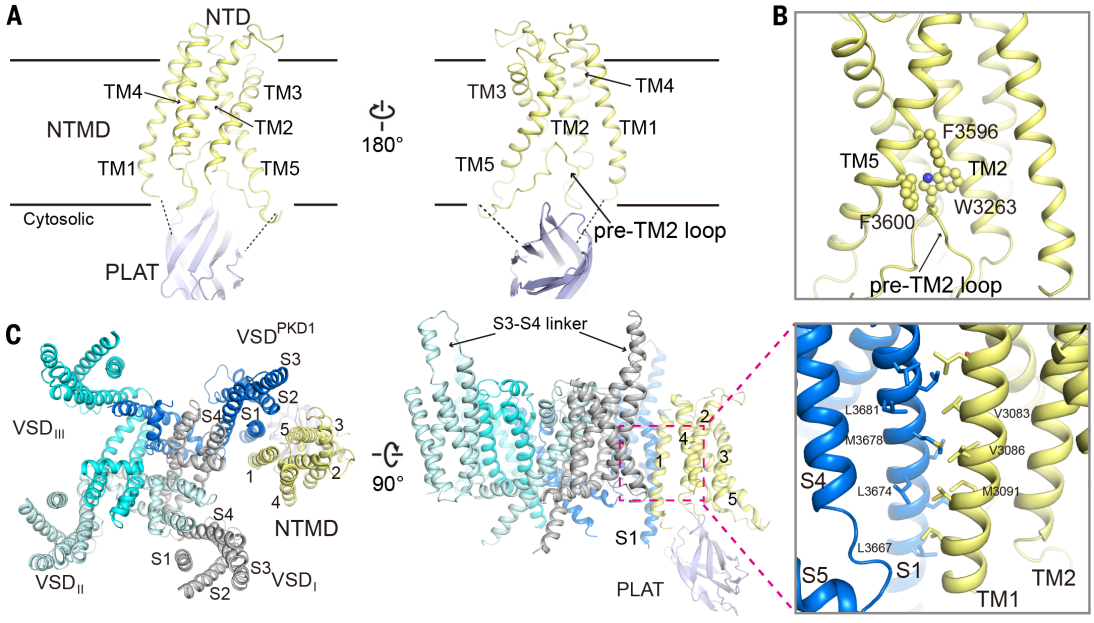


Fig. 4. Structure of PKD1-NTD. (**A**) The NTD comprises the NTMD and the cytosolic PLAT domain. Two opposing side views are shown. (**B**) The pre-TM2 loop appears to stabilize the structure of NTMD. The highly conserved Trp^{3263} on the pre-TM2 loop interacts with Phe^{3596} and Phe^{3600} on TM5. The discussed residues are shown as spheres. (**C**) The interactions between NTD and VSD^{PKD1}. An extracellular view (left) and a side view (right) are shown. The magnified view shows the interface between VSD^{PKD1} and NTD, which is constituted by hydrophobic residues, shown as sticks, on the S1 and TM1 segments.

Asp^{624} on the linker of S6 and PH2 in PKD2 I, and the other is the canonical VGIC contacts constituted by hydrophobic residues in VSD^{PKD1} and PD_I (Fig. 5, C to E). The interfaces between adjacent PKD2 subunits in the hetero-oligomer are nearly identical to those in the homotetrameric PKD2 or PKD2L1 (fig. S8C) (*26*).

Discussion

Structural determination of the PKD1-PKD2 complex affords the opportunity to map a large number of disease mutations (Fig. 6). Patients with mutations in PKD1 usually show more severe symptoms than those with mutations in PKD2 (*2*). The pathological mutations are unevenly distributed in PKD1, with the NTD and TOP domains being two hotspots (Fig. 6, left). The function of the NTD, a domain that is missing in PKD2 or other TRP channels, remains largely unknown except for limited characterizations of several disease mutations. Mutation R3277C in the NTMD has been suggested to affect the folding, glycosylation, and trafficking of PKD1 (*46*). Mutation R3162C in PLAT or W414G in TOP^{PKD2} abrogates proper trafficking of PKD1 or PKD2, respectively (*25*, *31*). A few disease-related residues are mapped to the hydrophobic interior. Substitution of these hydrophobic residues with charged ones—such as C3081R, W3726R, and L3834R—may destroy the structural integrity of NTMD and TOP^{PKD1} (*43*, *47*–*52*) (Fig. 6 and fig. S9A). Therefore, incorrect folding or trafficking of the PKD1-PKD2 complex can be pathogenic.

An unexpected observation is the lack of pathological mutations on the pore-forming segments in PKD1 (Fig. 6). Considering the distinct conformation of $S6^{PKD1}$, we compared the sequence of $S6^{PKD1}$ with that in different PKD1 and PKD2 homologs (fig. S9B). $S6^{PKD1}$ is the only one that contains multiple positively charged residues. In the present structure, these residues—Arg^{4100}, Arg^{4107}, and His^{4111}—protrude into the potential ion-conducting path, likely leading to poor, if any, ion permeability (*4*, *53*, *54*) (Fig. 3B and fig. S5D). By contrast, the S6 segments in PKD1L3, PKD2L1, and PKD2, which share sequence similarity, are enriched in hydrophobic residues, a common feature found in most TRP and VGIC channels (fig. S9B). The ion conduction of PKD2L1 and the hetero-oligomer between PKD1L3 and PKD2L1 can be recorded easily (*4*, *26*). Notably, several likely neutral mutations implicated in disease have been mapped to the $S6^{PKD1}$ or PD^{PKD1}, indicating that the pathogenic mechanism of PKD1 may be independent of a putative ion-conducting activity (*43*, *48*, *55*, *56*).

PKD1 and PKD2 are hypothesized to form a complex in the primary cilia (*11*, *57*). Recent electrophysiological characterizations on the renal collecting duct epithelium suggested that PKD2 could function as a monovalent cation-selective channel and that the ciliary ion conductance is independent of PKD1 (*53*, *58*). The structural features are consistent with the observation that PKD1 may be irrelevant to the ciliary ion current.

The near-atomic-resolution structure of the PKD1-PKD2 complex reveals the molecular details of the assembly of a hetero-oligomeric complex and provides the template for mapping a large number of disease mutations. Elucidation of the functional mechanism of PKD1 and PKD2 as well as the disease mechanism of the hundreds of ADPKD mutations await further investigations. Our structure serves as a framework for future biophysical, biochemical, cellular, and computational analysis of PKD1-PKD2 and ADPKD.

Materials and methods

Transient protein expression and purification

The codon-optimized full-length cDNAs for human PKD1 and PKD2 were synthesized by Tsingke Company (hPKD1 [Uniprot: P98161-1]; hPKD2 [Uniprot: Q13563-1])). For structural analysis, the truncated constructs of PKD1 (residues 3049 to 4169) with N-terminal triple FLAG tag (DYKDHDGDYKDHDIDYKDDDDK) and PKD2 (residues 185 to 723) with N-terminal Twin-Strep-Tag II (WSHPQFEKGGGSGGGSGGSAWSHPQFEK) from IBA GmbH were subcloned into the pCAG vector (*59*). For mass spectrometric (MS) analysis, the FLAG tag (DYKDDDDK) and Twin-Strep-tag II were fused at the C terminus of human full-length PKD1 and N terminus of human full-length PKD2, respectively (*60*, *61*). For immunofluorescence experiments, the FLAG tag is fused at its N terminus after the PKD1's signal peptide (MPPAAPARLALALGLGLWLGALA). The sequences of all constructs were verified before cell expression and protein purification.

The HEK 293F cells (Invitrogen) were cultured in SMM 293T-I medium (Sino Biological Inc.) at 37°C supplemented with 5% CO_2 in a Multitron-Pro shaker (Infors) at 130 rpm. When cell density reached 2×10^6 to 2.5×10^6 cells per ml, the cells were cotransfected with equal mass amount of the plasmids for PKD1and PKD2. For 1-liter HEK 293F cell culture, the two plasmids, each of ~1.5 mg, were premixed with 4-mg linear polyethylenimines (PEIs) (Polysciences) in 50-ml fresh medium for 15 to 30 min. The mixture was then added into cell culture followed by 15-min incubation. The transfected cells were cultured at 37°C for 24 hours and then at 30°C for an additional 48 hours.

The cells were harvested by centrifugation at 800*g* for 10 min and resuspended in the lysis buffer containing 20 mM HEPES, pH 7.5, 150 mM NaCl, 10% glycerol, 5 mM EDTA, and protease inhibitor cocktail (Amresco; 2 μg/ml aprotinin, 2 μg/ml leupeptin, 2 μg/ml pepstanin). The suspension was frozen in liquid nitrogen and stored at −80°C for further experiments.

For protein purification, the thawed suspension was supplemented with 1 mM PMSF prior

to homogenization. The lysate was incubated in the buffer containing 2% DDM (Anatrace), 0.5% soybean lipids (Sigma), and 0.4% CHS (Anatrace) at 4°C for 1.5 to 2 hours for membrane protein extraction. After ultracentrifugation at 18,700*g* for 40 to 60 min, the supernatant was collected and applied to the anti-FLAG M2 affinity gel (Sigma) at 4°C for three times. The resin was rinsed four times, each with 5 ml of buffer containing 20 mM HEPES, pH 7.5, 150 mM NaCl, 10% glycerol (w/v), 0.06% digitonin (w/v, Sigma), and the aforementioned protease inhibitor cocktail. The proteins were then eluted with wash buffer plus 300 to 400 μg/ml FLAG peptide. The eluent from the anti-FLAG M2 column was subsequently loaded to the Strep-Tactin resin (IBA) and incubated at 4°C for 1 hour. The resin was washed extensively by the same wash buffer before being eluted with wash buffer plus 5 mM D-Desthiobiotin (IBA). The protein eluent was concentrated by a 50-kDa cut-off Centricon (Millipore) and further purified by Superose-6 Increase column (GE Healthcare) in the buffer containing 20 mM HEPES, pH 7.5, 150 mM NaCl, and 0.1% digitonin. The peak fractions corresponding to the PKD1-PKD2 complex were pooled, concentrated, and supplemented with 5 mM EDTA. A typical final yield of the homogeneous complex through this procedure was approximately 2 to 3 μg per liter cell culture.

Cryo-EM data acquisition

Holey carbon grids (Quantifoil Au 300 mesh, R1.2/1.3) were glow-discharged in the Plasma Cleaner PDC-32G-2 (HARRICK PLASMA Company) with a vacuum for 2 min and mid force for 22 s. Aliquots (3 μl) of purified PKD1-PKD2 complex at concentration of ~10 mg/ml were placed on the glow-discharged grids, which were then blotted for 2.5 to 3.5 s and flash frozen in liquid ethane cooled by liquid nitrogen using Vitrobot Mark IV (Thermo Fisher Scientific) at 8°C and 100% humidity without wait time or blot force.

The grids were transferred to a Titan Krios TEM operated at 300 kV and equipped with Gatan GIF Quantum energy filter and Gatan K2 direct electron detector and Cs corrector. A total of 3761 zero-loss movie stacks were automatically collected using AutoEMation II (developed by J. Lei) in the super-resolution mode (*62*) with 20-eV slit in energy filter at a nominal magnification of 105,000× with defocus range from −1.0 to −2.0 μm. Each micrograph stack, which contained 32 frames, was exposed for 5.6 s with a total electron dose of ~50 $e^-/Å^2$. The stacks were motion corrected using MotionCor2 (*63*) with a binning factor of 2, resulting in a pixel size of 1.091 Å. Dose weighting was performed concurrently (*64*). The defocus values were estimated with Gctf (*65*).

Cryo-EM image processing

The procedure for image processing of PKD1-PKD2 complex is presented in fig. S3. A total of 902,194 particles were automatically picked with Gautomatch (developed by Kai Zhang, https://www.mrc-lmb.cam.ac.uk/kzhang/Gautomatch/). After 2D classification, 467,705 good particles were selected and subjected to 3D classification. The PKD2 map (EMDB ID: 8354) was low-pass filtered to 10 Å to be used as the initial model (*4*). All the particles were first subjected to global angular search 3D classification using RELION 2.0 with one class and step size of 7.5°. For each of the last several iterations of the global angular search 3D classification, a local angular search 3D classification was executed with class number being 8 or 15, step size of 3.75°, and local search range of 15°. A total of 210,935 good particles were combined and subjected to further 3D classification, from which 116,485 good particles were selected and subjected to 3D auto-refinement, resulting in a final resolution at 3.2 Å. To further improve the map quality for the flexible PKD2 III subunit, a total of 27,296 particles were selected by applying a mask during the "skip alignment" 3D classification, yielding a map with better S6a area with an overall resolution of 3.6 Å.

All resolutions mentioned above are determined according to the gold-standard Fourier shell correlation 0.143 criterion and with a high-resolution noise substitution method (*66*).

Model building and structure refinement

The PKD2 structure (PDB code: 5T4D) (*4*) was docked into the 3.2-Å map for model building of the majority of PKD1-CTD and PKD2 subunits in the complex. The VSD and TOP domains of PKD2 III, as well as the S5 and S6a of PKD1, were built based on the 3.6-Å map wherein these regions were better resolved. The secondary structural elements of PKD1-NTD were predicted in I-TASSER (*38*). Bulky residues (Phe, Trp, Arg, Lys, and Tyr) were used as reference for sequence assignment. Owing to the relatively poor resolutions, only poly-Ala chain was built for the PLAT domain taking account of the prediction from I-TASSER (*38*). The built model was real space refined using PHENIX with geometry restraints applied (*67*). The whole process was monitored to avoid overfitting by executing model refinements in two independent half maps, following the gold-standard refinement approach, and testing against each other (*67*). The final model was evaluated using MolProbity (*68*). The statistics

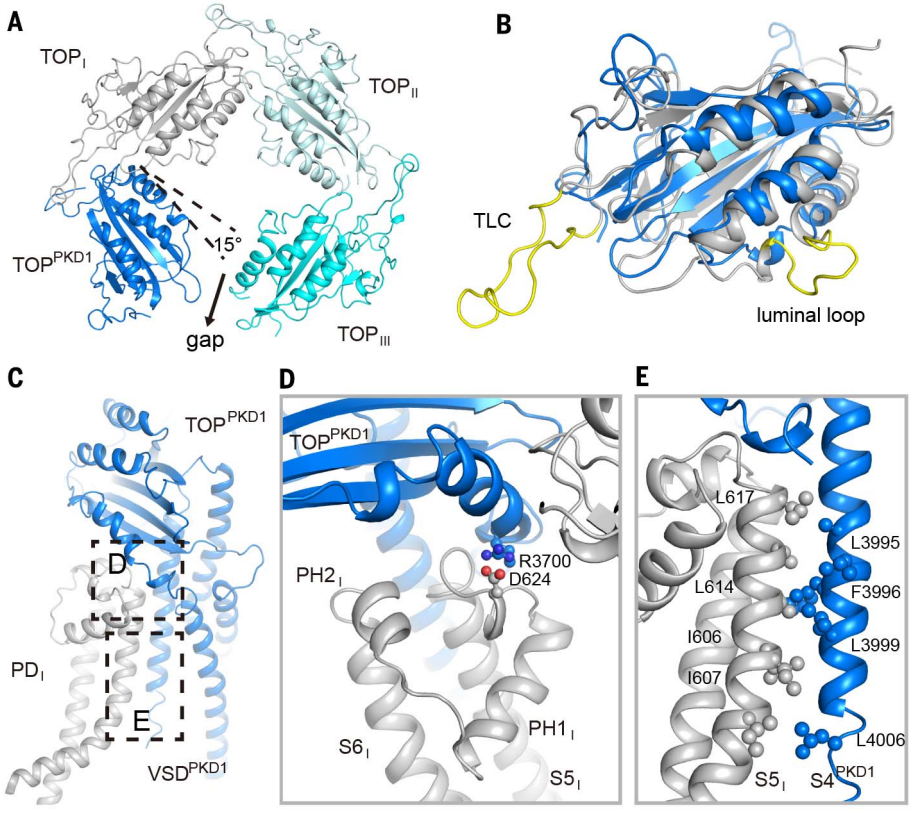


Fig. 5. Interactions between PKD1 and PKD2. (**A**) The ring of TOP domains in the heterotetramer is gapped owing to the deviation of TOP^{PKD1} from the fourfold symmetry. Shown is an extracellular view. (**B**) Structural comparison of TOP^{PKD1} and TOP^{PKD2}. The elements that mediate interactions among TOP domains in PKD2 (gray) but are missing (luminal loop) or invisible (TLC) in TOP^{PKD1} (blue) are colored yellow. (**C**) Two additional interfaces between PKD1 and PKD2. The indicated interfaces are illustrated in detail in (D) and (E). (**D**) The electrostatic interaction between TOP^{PKD1} and PD_I. (**E**) A hydrophobic interface between $S5_I$ and $S4^{PKD1}$. This interface observes the canonical domain-swapped folding principle of VGICs. The interface residues are shown as spheres.

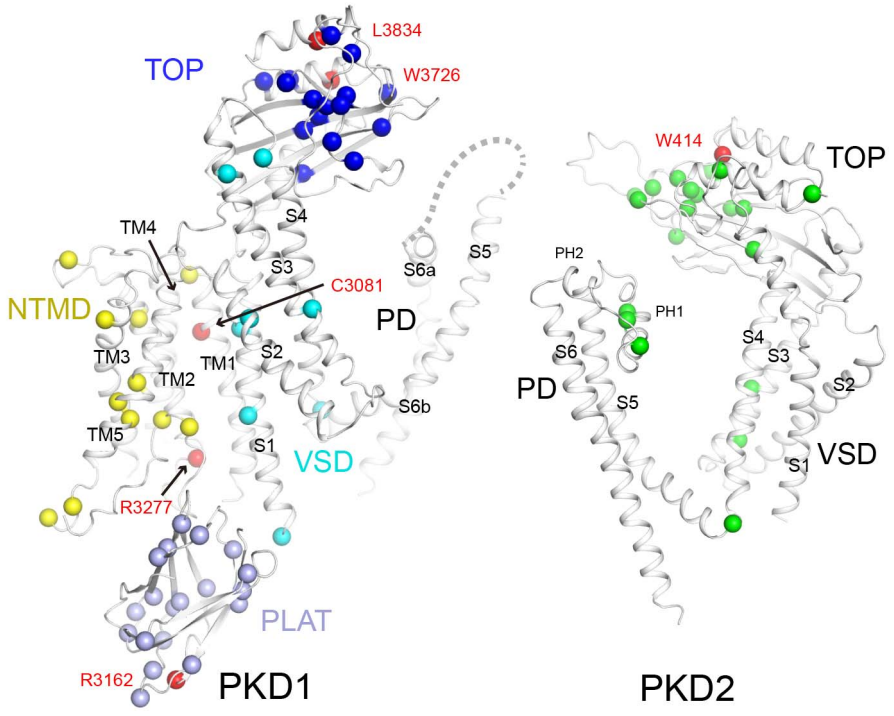


Fig. 6. Structural mapping of ADPKD mutations. The α carbon atoms of representative disease-related residues are shown as spheres and are domain colored. The missense mutations were summarized from the ADPKD Mutation Database (http://pkdb.mayo.edu/).

for data collection and model building and refinement are listed in table S1.

Quantitative proteomic analysis by label-free MS

Both full-length and truncated PKD1-PKD2 complexes were overexpressed and purified with the same protocol as mentioned above. Individual proteins were separated by SDS-PAGE followed by in-gel digestion for subsequent MS analysis. Briefly, all the proteins were treated with 25 mM dithiothreitol (DTT) to reduce disulfide bonds and then alkylated with 55 mM iodoacetamide. In-gel digestion was performed using sequencing grade-modified trypsin in 50 mM ammonium bicarbonate at 37°C overnight. The peptides were extracted twice with 1% trifluoroacetic acid in 50% acetonitrile aqueous solution for 30 min. The peptide extracts were then centrifuged in a SpeedVac to reduce the volume.

For LC-MS analysis, peptides were separated by a 60-min gradient elution at a flow rate of 0.300 μl/min with a Thermo-Dionex Ultimate 3000 HPLC system, which was directly interfaced with a Thermo LTQ-Orbitrap Velos pro mass spectrometer. The analytical column was a homemade fused silica capillary column (75 μm ID, 150 mm in length; Upchurch, Oak Harbor, WA) packed with C-18 resin (300 A, 5 μm; Varian, Lexington, MA). Mobile phase A contained 0.1% formic acid, and mobile phase B contained 100% acetonitrile supplemented with 0.1% formic acid. An LTQ-Orbitrap mass spectrometer was operated in the data-dependent acquisition mode using Xcalibur 2.0.7 software. A single full-scan mass spectrum in the Orbitrap (400 to 1800 m/z, 30,000 resolution) was followed by 20 data-dependent MS/MS scans in an ion trap at 35% normalized collision energy (CID).

The raw data were analyzed by MaxQuant (version 1.6.2.3) (*69*) using standard settings with the additional options to match between runs (between triplicates) with LFQ and iBAQ selected. The generated "proteingroups.txt" table was filtered for contaminants, reverse hits, number of unique peptides (>0) and number of peptides (>1) in Perseus (from MaxQuant package). To determine the stoichiometry of the target complexes, we compared the relative abundance of the identified interactors as measured by the iBAQ intensities (*35*). The sequences (PKD1 residues 3049 to 4169 and PKD2 residues 185 to 723) were used to search against both full-length and truncated samples to estimate the stoichiometry of the transmembrane domains. The data are listed in table S2.

Confocal microscopy immunofluorescence imaging

HEK 293T cells transfected with DNA combinations indicated in fig. S1 and Lipofectamine LTX with Plus (Invitrogen) were incubated for 24 hours. Cells were washed twice with PBS solution and then fixed with 4% paraformaldehyde in PBS for 15 min. After a 5-min wash with PBS for three times, the cells were either permeabilized with 0.5% Triton X-100 in TBS (25 mM Tris-HCl, pH 7.5, and 150 mM NaCl) for 20 min followed by three 5-min washes or processed without Triton X-100. The treated cells were then blocked with 4% bovine serum albumin (BSA) in TBS for 1 hour and incubated with the polyclonal anti-FLAG antibody in the same blocking solution at room temperature for 1 hour. After three 10-min washes with TBS, the cells were incubated with fluorescein (FITC)-conjugated goat anti-rabbit IgG in 4% BSA at room temperature for 1 hour. After another three 10-min washes with TBS, cells were mounted on slides and imaged with a confocal microscope.

The confocal fluorescence imaging experiments were performed with a ZEISS laser scanning confocal microscopy (LSM710). Data were collected and analyzed by ZEN 2012 Light Edition software.

REFERENCES AND NOTES

1. V. E. Torres, P. C. Harris, Y. Pirson, Autosomal dominant polycystic kidney disease. *Lancet* **369**, 1287–1301 (2007). doi: 10.1016/S0140-6736(07)60601-1; pmid: 17434405
2. P. C. Harris, V. E. Torres, Polycystic kidney disease. *Annu. Rev. Med.* **60**, 321–337 (2009). doi: 10.1146/annurev.med.60.101707.125712; pmid: 18947299
3. P. Igarashi, S. Somlo, Polycystic kidney disease. *J. Am. Soc. Nephrol.* **18**, 1371–1373 (2007). doi: 10.1681/ASN.2007030299; pmid: 17429047
4. P. S. Shen *et al.*, The structure of the polycystic kidney disease channel PKD2 in lipid nanodiscs. *Cell* **167**, 763–773.e11 (2016). doi: 10.1016/j.cell.2016.09.048; pmid: 27768895
5. W. Zheng *et al.*, Hydrophobic pore gates regulate ion permeation in polycystic kidney disease 2 and 2L1 channels. *Nat. Commun.* **9**, 2302 (2018). doi: 10.1038/s41467-018-04586-x; pmid: 29899465
6. J. Zhou, Polycystins and primary cilia: Primers for cell cycle progression. *Annu. Rev. Physiol.* **71**, 83–113 (2009). doi: 10.1146/annurev.physiol.70.113006.100621; pmid: 19572811
7. J. Hughes *et al.*, The polycystic kidney disease 1 (*PKD1*) gene encodes a novel protein with multiple cell recognition domains. *Nat. Genet.* **10**, 151–160 (1995). doi: 10.1038/ng0695-151; pmid: 7663510
8. T. Mochizuki *et al.*, *PKD2*, a gene for polycystic kidney disease that encodes an integral membrane protein. *Science* **272**, 1339–1342 (1996). doi: 10.1126/science.272.5266.1339; pmid: 8650545
9. A. Patel, E. Honoré, Polycystins and renovascular mechanosensory transduction. *Nat. Rev. Nephrol.* **6**, 530–538 (2010). doi: 10.1038/nrneph.2010.97; pmid: 20625375
10. S. Kim *et al.*, The polycystin complex mediates Wnt/Ca^{2+} signalling. *Nat. Cell Biol.* **18**, 752–764 (2016). doi: 10.1038/ncb3363; pmid: 27214281
11. S. M. Nauli *et al.*, Polycystins 1 and 2 mediate mechanosensation in the primary cilium of kidney cells. *Nat. Genet.* **33**, 129–137 (2003). doi: 10.1038/ng1076; pmid: 12514735
12. V. Chauvet *et al.*, Mechanical stimuli induce cleavage and nuclear translocation of the polycystin-1 C terminus. *J. Clin. Invest.* **114**, 1433–1443 (2004). doi: 10.1172/JCI21753; pmid: 15545994
13. Y. Yang, B. E. Ehrlich, Structural studies of the C-terminal tail of polycystin-2 (PC2) reveal insights into the mechanisms used for the functional regulation of PC2. *J. Physiol.* **594**, 4141–4149 (2016). doi: 10.1113/JP270933; pmid: 26857659
14. L. Geng *et al.*, Distribution and developmentally regulated expression of murine polycystin. *Am. J. Physiol.* **272**, F451–F459 (1997). pmid: 9140045
15. C. J. Ward *et al.*, Polycystin, the polycystic kidney disease 1 protein, is expressed by epithelial cells in fetal, adult, and polycystic kidney. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 1524–1528 (1996). doi: 10.1073/pnas.93.4.1524; pmid: 8643665
16. L. Foggensteiner *et al.*, Cellular and subcellular distribution of polycystin-2, the protein product of the *PKD2* gene. *J. Am. Soc. Nephrol.* **11**, 814–827 (2000). pmid: 10770959
17. K. Hanaoka *et al.*, Co-assembly of polycystin-1 and -2 produces unique cation-permeable currents. *Nature* **408**, 990–994 (2000). doi: 10.1038/35050128; pmid: 11140688
18. M. Delling *et al.*, Primary cilia are not calcium-responsive mechanosensors. *Nature* **531**, 656–660 (2016). doi: 10.1038/nature17426; pmid: 27007841

19. F. Qian *et al*., PKD1 interacts with PKD2 through a probable coiled-coil domain. *Nat. Genet.* **16**, 179–183 (1997). doi: 10.1038/ng0697-179; pmid: 9171830
20. L. Tsiokas, E. Kim, T. Arnould, V. P. Sukhatme, G. Walz, Homo- and heterodimeric interactions between the gene products of *PKD1* and *PKD2*. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 6965–6970 (1997). doi: 10.1073/pnas.94.13.6965; pmid: 9192675
21. Y. Yu *et al*., Structural and molecular basis of the assembly of the TRPP2/PKD1 complex. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 11558–11563 (2009). doi: 10.1073/pnas.0903684106; pmid: 19556541
22. J. Zhu *et al*., Structural model of the TRPP2/PKD1 C-terminal coiled-coil complex produced by a combined computational and experimental approach. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 10133–10138 (2011). doi: 10.1073/pnas.1017669108; pmid: 21642537
23. V. Babich *et al*., The N-terminal extracellular domain is required for polycystin-1-dependent channel activity. *J. Biol. Chem.* **279**, 25582–25589 (2004). doi: 10.1074/jbc.M402829200; pmid: 15060061
24. S. Feng *et al*., Identification and functional characterization of an N-terminal oligomerization domain for polycystin-2. *J. Biol. Chem.* **283**, 28471–28479 (2008). doi: 10.1074/jbc.M803834200; pmid: 18701462
25. Y. Cai *et al*., Altered trafficking and stability of polycystins underlie polycystic kidney disease. *J. Clin. Invest.* **124**, 5129–5144 (2014). doi: 10.1172/JCI67273; pmid: 25365220
26. Q. Su *et al*., Cryo-EM structure of the polycystic kidney disease-like channel PKD2L1. *Nat. Commun.* **9**, 1192 (2018). doi: 10.1038/s41467-018-03606-0; pmid: 29567962
27. M. Grieben *et al*., Structure of the polycystic kidney disease TRP channel polycystin-2 (PC2). *Nat. Struct. Mol. Biol.* **24**, 114–122 (2017). doi: 10.1038/nsmb.3343; pmid: 27991905
28. M. Wilkes *et al*., Molecular insights into lipid-assisted Ca^{2+} regulation of the TRP channel polycystin-2. *Nat. Struct. Mol. Biol.* **24**, 123–130 (2017). doi: 10.1038/nsmb.3357; pmid: 28092368
29. K. Venkatachalam, C. Montell, TRP channels. *Annu. Rev. Biochem.* **76**, 387–417 (2007). doi: 10.1146/annurev.biochem.75.103004.142819; pmid: 17579562
30. M. Bycroft *et al*., The structure of a PKD domain from polycystin-1: Implications for polycystic kidney disease. *EMBO J.* **18**, 297–305 (1999). doi: 10.1093/emboj/18.2.297; pmid: 9889186
31. Y. Xu *et al*., The polycystin-1, lipoxygenase, and α-toxin domain regulates polycystin-1 trafficking. *J. Am. Soc. Nephrol.* **27**, 1159–1173 (2016). doi: 10.1681/ASN.2014111074; pmid: 26311459
32. A. Bateman, R. Sandford, The PLAT domain: A new piece in the PKD1 puzzle. *Curr. Biol.* **9**, R588–R590 (1999). doi: 10.1016/S0960-9822(99)80380-7; pmid: 10469604
33. D. H. Grimm *et al*., Polycystin-1 distribution is modulated by polycystin-2 expression in mammalian cells. *J. Biol. Chem.* **278**, 36786–36793 (2003). doi: 10.1074/jbc.M306536200; pmid: 12840011
34. P. B. Rosenthal, R. Henderson, Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003). doi: 10.1016/j.jmb.2003.07.013; pmid: 14568533
35. A. H. Smits, P. W. Jansen, I. Poser, A. A. Hyman, M. Vermeulen, Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics. *Nucleic Acids Res.* **41**, e28 (2013). doi: 10.1093/nar/gks941; pmid: 23066101
36. R. van Nuland *et al*., Quantitative dissection and stoichiometry determination of the human SET1/MLL histone methyltransferase complexes. *Mol. Cell. Biol.* **33**, 2067–2077 (2013). doi: 10.1128/MCB.01742-12; pmid: 23508102
37. C. Smaczniak *et al*., Proteomics-based identification of low-abundance signaling and regulatory protein complexes in native plant tissues. *Nat. Protoc.* **7**, 2144–2158 (2012). doi: 10.1038/nprot.2012.129; pmid: 23196971
38. J. Yang *et al*., The I-TASSER Suite: Protein structure and function prediction. *Nat. Methods* **12**, 7–8 (2015). doi: 10.1038/nmeth.3213; pmid: 25549265
39. U. A. Hellmich, R. Gaudet, Structural biology of TRP channels. *Handb. Exp. Pharmacol.* **223**, 963–990 (2014). doi: 10.1007/978-3-319-05161-1_10; pmid: 24961976
40. W. A. Catterall, Forty years of sodium channels: Structure, function, pharmacology, and epilepsy. *Neurochem. Res.* **42**, 2495–2504 (2017). doi: 10.1007/s11064-017-2314-9; pmid: 28589518
41. E. Grandi *et al*., Potassium channels in the heart: Structure, function and regulation. *J. Physiol.* **595**, 2209–2228 (2017). doi: 10.1113/JP272864; pmid: 27861921
42. L. Holm, L. M. Laakso, Dali server update. *Nucleic Acids Res.* **44**, W351–W355 (2016). doi: 10.1093/nar/gkw357pmid: 27131377
43. S. Rossetti *et al*., Identification of gene mutations in autosomal dominant polycystic kidney disease through targeted resequencing. *J. Am. Soc. Nephrol.* **23**, 915–933 (2012). doi: 10.1681/ASN.2011101032; pmid: 22383692
44. X. Zhou *et al*., Cryo-EM structures of the human endolysosomal TRPML3 channel in three distinct states. *Nat. Struct. Mol. Biol.* **24**, 1146–1154 (2017). doi: 10.1038/nsmb.3502; pmid: 29106414
45. Z. Salehi-Najafabadi *et al*., Extracellular loops are essential for the assembly and function of polycystin receptor-ion channel complexes. *J. Biol. Chem.* **292**, 4210–4221 (2017). doi: 10.1074/jbc.M116.767897; pmid: 28154010
46. K. Hopp *et al*., Functional polycystin-1 dosage governs autosomal dominant polycystic kidney disease severity. *J. Clin. Invest.* **122**, 4257–4273 (2012). doi: 10.1172/JCI64313; pmid: 23064367
47. J. Hoefele, K. Mayer, M. Scholz, H. G. Klein, Novel *PKD1* and *PKD2* mutations in autosomal dominant polycystic kidney disease (ADPKD). *Nephrol. Dial. Transplant.* **26**, 2181–2188 (2011). doi: 10.1093/ndt/gfq720; pmid: 21115670
48. S. Rossetti *et al*., Comprehensive molecular diagnostics in autosomal dominant polycystic kidney disease. *J. Am. Soc. Nephrol.* **18**, 2143–2160 (2007). doi: 10.1681/ASN.2006121387; pmid: 17582161
49. G. Aguiari *et al*., Novel splicing and missense mutations in autosomal dominant polycystic kidney disease 1 (*PKD1*) gene: Expression of mutated genes. *Hum. Mutat.* **16**, 444–445 (2000). doi: 10.1002/1098-1004(200011)16:5<444:AID-HUMU11>3.0.CO;2-C; pmid: 11058904
50. B. Reed *et al*., Presence of de novo mutations in autosomal dominant polycystic kidney disease patients without family history. *Am. J. Kidney Dis.* **52**, 1042–1050 (2008). doi: 10.1053/j.ajkd.2008.05.015; pmid: 18640754
51. S. Rossetti *et al*., A complete mutation screen of the ADPKD genes by DHPLC. *Kidney Int.* **61**, 1588–1599 (2002). doi: 10.1046/j.1523-1755.2002.00326.x; pmid: 11967008
52. Y. C. Tan *et al*., Novel method for genomic analysis of *PKD1* and *PKD2* mutations in autosomal dominant polycystic kidney disease. *Hum. Mutat.* **30**, 264–273 (2009). doi: 10.1002/humu.20842; pmid: 18837007
53. X. Liu *et al*., Polycystin-2 is an essential ion channel subunit in the primary cilium of the renal collecting duct epithelium. *eLife* **7**, e33183 (2018). doi: 10.7554/eLife.33183; pmid: 29443690
54. Y. Yu *et al*., Molecular mechanism of the assembly of an acid-sensing receptor ion channel complex. *Nat. Commun.* **3**, 1252 (2012). doi: 10.1038/ncomms2257; pmid: 2321238
55. S. Bataille, Y. Berland, M. Fontes, S. Burtey, High resolution melt analysis for mutation screening in *PKD1* and *PKD2*. *BMC Nephrol.* **12**, 57 (2011). doi: 10.1186/1471-2369-12-57; pmid: 22008521
56. L. Ding *et al*., Novel mutations of *PKD1* gene in Chinese patients with autosomal dominant polycystic kidney disease. *Nephrol. Dial. Transplant.* **17**, 75–80 (2002). doi: 10.1093/ndt/17.1.75; pmid: 11773467
57. M. Ma, A. R. Gallagher, S. Somlo, Ciliary mechanisms of cyst formation in polycystic kidney disease. *Cold Spring Harb. Perspect. Biol.* **9**, a028209 (2017). doi: 10.1101/cshperspect.a028209; pmid: 28320755
58. S. J. Kleene, N. K. Kleene, The native TRPP2-dependent channel of murine renal primary cilia. *Am. J. Physiol. Renal Physiol.* **312**, F96–F108 (2017). doi: 10.1152/ajprenal.00272.2016; pmid: 27760766
59. T. Matsuda, C. L. Cepko, Electroporation and RNA interference in the rodent retina in vivo and in vitro. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 16–22 (2004). doi: 10.1073/pnas.2235688100; pmid: 14603031
60. C. J. Gloeckner, K. Boldt, A. Schumacher, M. Ueffing, Tandem affinity purification of protein complexes from mammalian cells by the Strep/FLAG (SF)-TAP tag. *Methods Mol. Biol.* **564**, 359–372 (2009). doi: 10.1007/978-1-60761-157-8_21; pmid: 19544034
61. C. J. Gloeckner, K. Boldt, M. Ueffing, Strep/FLAG tandem affinity purification (SF-TAP) to study protein interactions. *Curr. Protoc. Protein Sci.* **57**, 19.20.1–19.20.19 (2009). doi: 10.1007/978-1-60761-157-8_21; pmid: 19544034
62. J. Lei, J. Frank, Automated acquisition of cryo-electron micrographs for single particle reconstruction on an FEI Tecnai electron microscope. *J. Struct. Biol.* **150**, 69–80 (2005). doi: 10.1016/j.jsb.2005.01.002; pmid: 15797731
63. S. Q. Zheng *et al*., MotionCor2: Anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017). pmid: 28250466
64. T. Grant, N. Grigorieff, Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. *eLife* **4**, e06980 (2015). doi: 10.7554/eLife.06980; pmid: 26023829
65. K. Zhang, Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016). doi: 10.1016/j.jsb.2015.11.003; pmid: 26592709
66. S. Chen *et al*., High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. *Ultramicroscopy* **135**, 24–35 (2013). doi: 10.1016/j.ultramic.2013.06.004; pmid: 23872039
67. P. D. Adams *et al*., PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010). pmid: 20124702
68. V. B. Chen *et al*., MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 12–21 (2010). doi: 10.1107/S0907444909042073; pmid: 20057044
69. J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008). doi: 10.1038/nbt.1511; pmid: 19029910
70. N. Alexander, N. Woetzel, J. Meiler, bcl:Cluster: A method for clustering biological molecules coupled with visualization in the Pymol Molecular Graphics System. *IEEE Int. Conf. Comput. Adv. Bio Med. Sci.* **2011**, 13–18 (2011). pmid: 27818847

ACKNOWLEDGMENTS

We thank C. Yan and J. Zeng for advice on model building; Xiaomin Li, Xiaomei Li, and X. Hu for technical support during cryo-EM data collection; H. Deng and X. Meng at the Center of Biomedical Analysis, Tsinghua University, for MS analysis; and C. Huang and S. Gao at Peking University for advice on MS analysis. We thank the Tsinghua University Branch of China National Center for Protein Sciences (Beijing) for providing the facility support. The computation was completed on the "Explorer 100" cluster system of Tsinghua National Laboratory for Information Science and Technology. **Funding:** This work was supported by funds from the National Natural Science Foundation of China (31621092, 31430020, 81370784, and 81770659) and the National Key R&D Program (grant 2016YFA0501100) from the Ministry of Science and Technology of China. **Author contributions:** Y.S., C.M., S.Y., and T.W. conceived the project. Y.S., Q.S., and F.H. designed experiments. Q.S., F.H., X.G., and J.L. performed experiments. All authors contributed to data analysis. Y.S., Q.S., and F.H. wrote the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The cryo-EM maps and the structure have been deposited to the Electron Microscopy Data Bank (EMDB 6991 and EMDB 6992) and the Protein Data Bank (PDB 6A70), respectively.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6406/eaat9819/suppl/DC1
Figs. S1 to S9
Tables S1 to S3
References (*71*, *72*)

24 April 2018; accepted 30 July 2018
Published online 9 August 2018
10.1126/science.aat9819

RESEARCH ARTICLE

TOPOLOGICAL OPTICS

Optical skyrmion lattice in evanescent electromagnetic fields

S. Tsesses[1], E. Ostrovsky[1], K. Cohen[1], B. Gjonaj[2], N. H. Lindner[3], G. Bartal[1]*

Topological defects play a key role in a variety of physical systems, ranging from high-energy to solid-state physics. A skyrmion is a type of topological defect that has shown promise for applications in the fields of magnetic storage and spintronics. We show that optical skyrmion lattices can be generated using evanescent electromagnetic fields and demonstrate this using surface plasmon polaritons, imaged by phase-resolved near-field optical microscopy. We show how the optical skyrmion lattice exhibits robustness to imperfections while the topological domain walls in the lattice can be continuously tuned, changing the spatial structure of the skyrmions from bubble type to Néel type. Extending the generation of skyrmions to photonic systems provides various possibilities for applications in optical information processing, transfer, and storage.

Topological defects are field configurations that cannot be deformed to a standard, smooth shape. They are at the core of many fascinating phenomena in hydrodynamics, aerodynamics, exotic phases of matter (*1*, *2*), cosmology (*3*), and optics (*4*) and, in many cases, are of importance to practical applications. The intricate dynamics of a multitude of topological defects and the efforts to control them are of key importance in high-temperature superconductivity (*5*) and topological phase transitions such as the Berezinskii-Kosterlitz-Thouless transition (*6*).

One type of topological defect is a skyrmion (*7*), a topologically stable configuration of a three-component vector field in two dimensions. Skyrmions were initially developed theoretically in elementary particles and have since been demonstrated in Bose-Einstein condensates (*8*) and nematic liquid crystals (*9*) and as a phase transition in chiral magnets (*10*, *11*). The skyrmion lattice phase and single magnetic skyrmions (*12*, *13*) are considered a promising route toward high-density magnetic information storage and transfer (*14*, *15*), as they are very robust to material defects and can be driven by low applied currents (*12*, *16*, *17*). A skyrmion may take on various shapes, which are all topologically equivalent. Bloch-type (*11*) and Néel-type (*18*) skyrmions exhibit a smoothly varying field configuration, with the derivatives of the vector field spread out in space. In bubble-type skyrmions (*19*, *20*), the variations of the vector field are confined to line-like areas, known as topological domain walls, which separate two domains in which the field vectors are opposite.

In optics, topological phenomena have been researched intensely in the past decade. Since the first observation of photonic topological insulators (*21*), optical topological phenomena have been rigorously studied, both theoretically and experimentally (*22*–*25*), with applications such as topologically protected lasing (*26*–*28*). Indeed, topological defects in optics were first extensively explored via phase and polarization singularities, both in free-space propagating light (*29*–*31*) and in two-dimensionally confined light (*32*–*34*). However, only recently has there been any experimental investigation of optical topological domain walls (*35*), whereas the field of optical skyrmions has so far remained untapped.

Formation of optical skyrmion lattices

The configuration of a three-component real vector field on a two-dimensional (2D) space provides a smooth mapping of that space to the unit sphere. The topological invariant that identifies skyrmions counts the number of times that the field configuration covers the entire sphere. This topological invariant, which we denote by S, is known as the skyrmion number and takes integer values. For the skyrmion to be topologically robust, the space on which it is defined cannot have a boundary. This condition is indeed satisfied by a periodic field configuration, in which a skyrmion is obtained in each unit cell of a lattice. For such a "skyrmion lattice" configuration, the skyrmion number can be written in an integral form

$$S = \frac{1}{4\pi}\int_A s\,\mathrm{d}A \qquad (1)$$

where the area A covers one unit cell of the lattice and $s = \vec{\mathbf{e}} \cdot [(\partial\vec{\mathbf{e}}/\partial x) \times (\partial\vec{\mathbf{e}}/\partial y)]$ is the skyrmion number density; $\vec{\mathbf{e}}$ is a real, normalized, three-component field; and x and y are directions

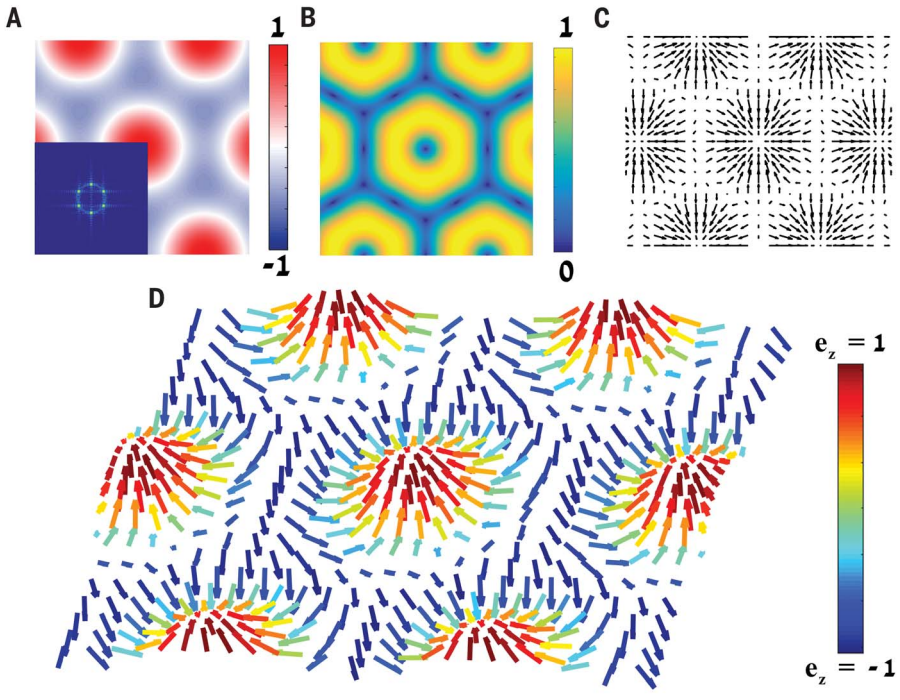


Fig. 1. Calculated electric field distribution of the optical skyrmion lattice. (**A**) Axial (out-of-plane) electric field, according to Eq. 2, with the Fourier decomposition shown in the inset. The color scale indicates the value of $|E_z^{(\omega)}|/E_0$. (**B**) Amplitude of the transverse (in-plane) electric field (Eq. 3). The color scale indicates the value of $\sqrt{|E_x^{(\omega)}|^2 + |E_y^{(\omega)}|^2}/\sqrt{2}E_0$. (**C**) Vector representation of the transverse electric field, showing polarization singularities at the center of each lattice site. The arrows indicate the magnitude and orientation of the in-plane electric field. (**D**) Vector representation of the local unit vector of the electric field (color coded for the value of its axial component), showing that each lattice site is a Néel-type skyrmion. The arrows indicate the direction of the local unit vector of the electric field.

[1]Andrew and Erna Viterbi Department of Electrical Engineering, Technion–Israel Institute of Technology, 3200003 Haifa, Israel. [2]Faculty of Medical Sciences, Albanian University, Durrës Street, Tirana 1000, Albania. [3]Physics Department, Technion–Israel Institute of Technology, 3200003 Haifa, Israel.
*Corresponding author. Email: guy@ee.technion.ac.il

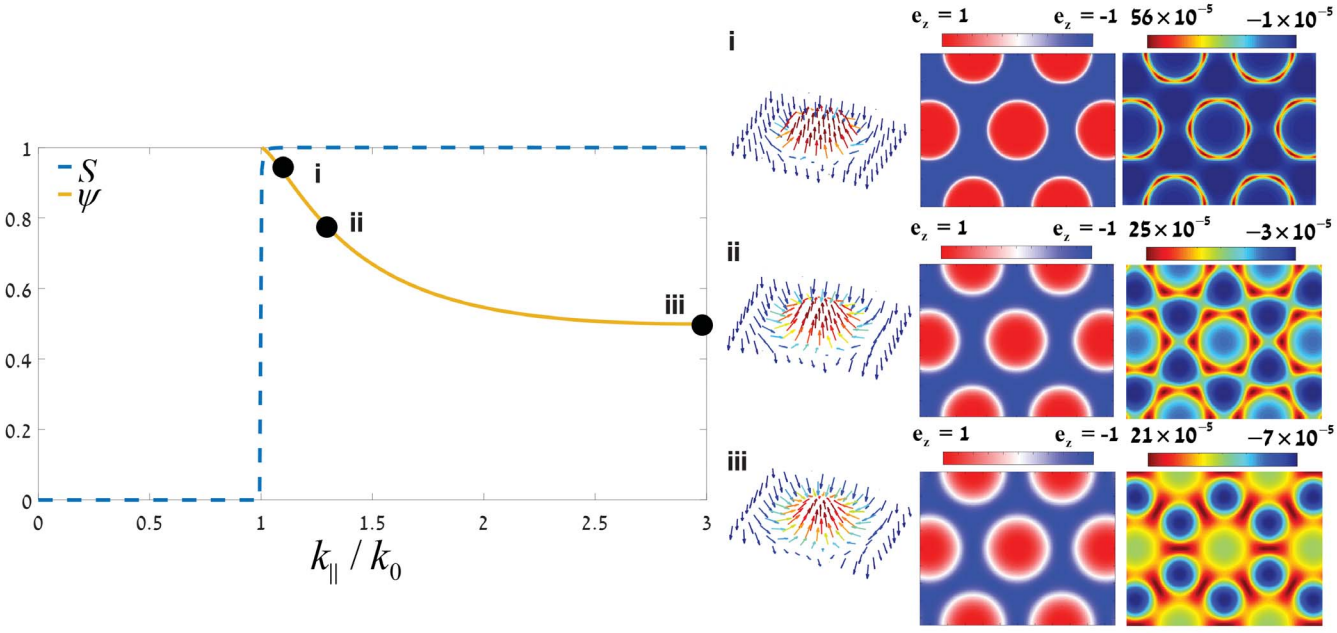


Fig. 2. Tuning of optical topological domain walls, from bubble-type to Néel-type skyrmions. (Left) The calculated skyrmion number (dashed blue line) and the skyrmion number density contrast (solid gold line) as a function of the transverse wave vector of the electromagnetic waves. Once the electromagnetic waves gain evanescence in free space, the skyrmion number jumps to the value of 1, independent of the change in the transverse wave vector. Points i to iii have a different skyrmion number density contrast, quantifying the transition from bubble-type to Néel-type skyrmions. (Right) The characteristics of a single lattice site at each of these points. Shown, from left to right, are the local unit vector of the electric field (color coded in the same manner as in Fig. 1), its axial (out-of-plane) component, and the skyrmion number density. A clear shift from very narrow domain walls separating between two field states (point i) to smeared domain walls (point ii) and, finally, to virtually nonlocal domain walls (point iii) arises by increasing the transverse wave vector.

in the 2D plane. The skyrmion number S, being an integer, is robust to deformations of the field $\vec{\mathbf{e}}$ as long as $\vec{\mathbf{e}}$ remains nonsingular and maintains the periodicity of the lattice.

We find that the electric field vector of electromagnetic waves can be structured so as to meet all the requirements needed to create a skyrmion lattice, similar to the magnetization vector of the skyrmion lattices in chiral magnets. Being a 3D field in a 2D space, optical skyrmions must be formed by electromagnetic waves confined to two dimensions, such as in guided waves. A more thorough explanation as to why free-space electromagnetic waves do not exhibit skyrmions can be found in the supplementary text.

Consider an electric field comprised of six interfering transverse-magnetic (TM) guided waves with equal amplitudes, freely propagating in the transverse (x-y) plane and evanescently decaying in the axial direction (z). The six waves are directed toward each other in the transverse plane and possess transverse wave vectors of similar magnitude, such that they create three standing waves oriented about 0°, 60°, and 120°. The axial (out-of-plane) field component in the frequency (ω) domain can therefore be expressed as a sum of three cosine functions

$$E_z^{(\omega)} = E_0 e^{-|k_z|z} \sum_{\theta=-\frac{\pi}{3},0,\frac{\pi}{3}} \cos\{k_{\|}[\cos(\theta)x + \sin(\theta)y]\} \quad (2)$$

where E_0 is a real normalization constant and $k_{\|}$ and k_z are the transverse and axial components of the wave vector ($k_{\|}$ is real and k_z is imaginary), such that $k_{\|}^2 + k_z^2 = k_0^2$ (k_0 is the free-space wave number). The transverse (in-plane) electric field components in the frequency domain can be readily derived from Maxwell's equations:

$$\begin{bmatrix} E_x^{(\omega)} \\ E_y^{(\omega)} \end{bmatrix} = -E_0 \frac{|k_z|}{k_{\|}} e^{-|k_z|z} \sum_{\theta=-\frac{\pi}{3},0,\frac{\pi}{3}} \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} \sin\{k_{\|}[\cos(\theta)x + \sin(\theta)y]\} \quad (3)$$

Namely, for waves evanescently decaying in one dimension, all electric field components are entirely real (up to a global phase), allowing the definition of a real unit vector $\vec{\mathbf{e}} = \vec{E}/|\vec{E}|$ associated with the electric field and enabling a description of the field configuration as an optical skyrmion lattice. This is a direct consequence of the phase added in the spin-momentum locking process of evanescent electromagnetic fields (*36*) and hence does not hold for propagating waves [see (*37*), section S.1].

Figure 1 depicts the electric field described by Eqs. 2 and 3 for $|k_z| \approx |k_{\|}| > k_0$. The axial electric field has the form of a hexagonally symmetric lattice (Fig. 1A). The transverse field follows the same symmetry yet possesses pronounced polarization singularities, which are expressed by zero-amplitude points at the center of each lattice site (Fig. 1B), at which the field direction is ill-defined (Fig. 1C). The normalized 3D electric field (Fig. 1D) confirms the formation of a skyrmion lattice; each lattice site exhibits the distinct features of a Néel-type skyrmion (*18*), and the calculated skyrmion number of each site, using Eq. 1, is $S = 1$.

Formation of the optical skyrmion lattice can be represented in the momentum space as a transition from free-space propagation in 3D ($|k_{\|}| < k_0$) to guided mode propagation in 2D ($|k_{\|}| > k_0$). To rigorously describe this transition, we must expand the skyrmion number definition to complex electromagnetic fields by defining $\vec{\mathbf{e}}$ in Eq. 1 as the real part of the local unit vector of the electric field. This new definition is consistent with the prior description of optical skyrmions in the ideal case, which is determined by Eqs. 2 and 3 in a lossless medium. Figure 2 presents the transition by showing the skyrmion number in a single lattice site as a function of $k_{\|}$ using the above definition for $\vec{\mathbf{e}}$.

Figure 2 also shows the skyrmion number density contrast [$\psi = (s_{max} + s_{min})/(s_{max} - s_{min})$] in a single lattice site as a function of $k_{\|}$, which is a parameter providing a quantitative measure of the spatial confinement of the skyrmion density. For $k_{\|}$ slightly larger than k_0, the skyrmions are spatially confined (density contrast close to 1), with clear domain walls separating between two specific field states, effectively creating bubble-type skyrmions (point i). As $k_{\|}$ grows, the domain walls start to smear (point ii), creating skyrmions with increasingly uniform skyrmion number density and converging to the Néel-type field formation shown in Fig. 1 (point iii in Fig. 2, density contrast of 0.5). This stems directly from the scaling factor $|k_z|/k_{\|}$, which can be tuned by varying the effective index (normalized propagation constant) of a guided mode.

Observation of an optical skyrmion lattice

Observing the optical skyrmion lattice has several prerequisites: First, it requires a physical

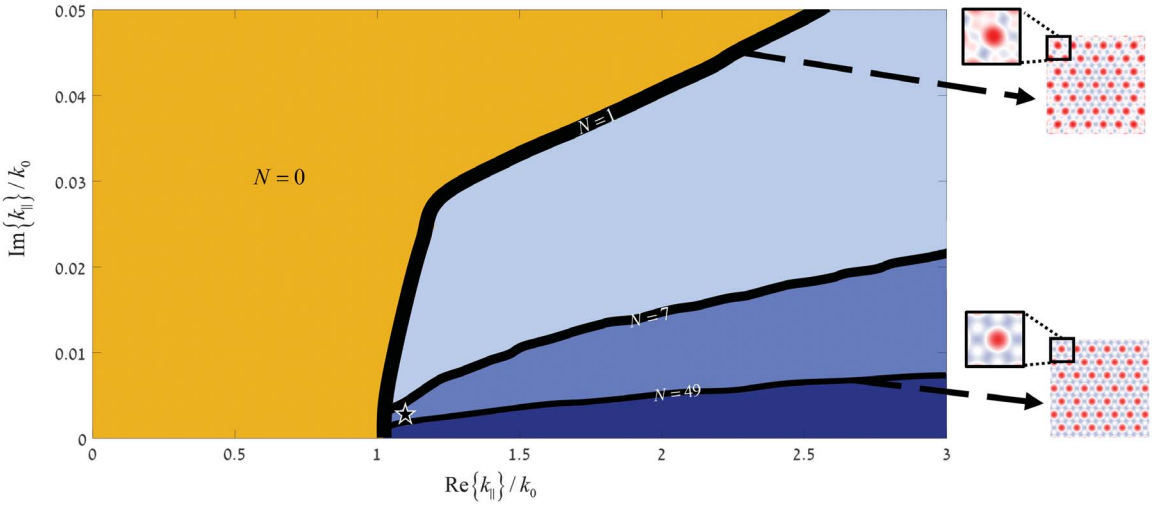


Fig. 3. Theoretical persistence of optical skyrmions in the presence of loss. A contour plot of the maximal number of well-defined skyrmions in a lattice (N), as a function of the real and imaginary parts of the transverse wave vector. The broad black lines represent the areas where the lattice consists of N = 1, 7, and 49 well-defined skyrmions. The amplitude distributions of an N = 1 lattice (top right) and an N = 49 lattice (bottom right) are shown, with close-up images showing the lattice distortion in noncentral sites. Clearly, in the N = 1 lattice, only the central lattice site is well defined. The star represents the experimental conditions of this work (N = 37). These results show that, by determining the number of well-defined skyrmions in a lattice, it is possible to determine the maximal amount of losses that will destroy its topological invariant. Moreover, these results show that even guided modes with a relatively high amount of loss (decaying after 100 oscillations) can create skyrmion lattices.

system that allows 2D guided waves to propagate in six specific directions while interfering them with carefully controlled phase differences. Second, it necessitates a measurement apparatus that will enable the phase-resolved imaging of such electromagnetic waves at a resolution far better than the optical diffraction limit. Additionally, a nonideal physical system—for example, one that is finite and with inherent losses—provides an excellent platform to examine the robustness of the topological properties of the optical skyrmions.

Such losses distort the unit vector $\vec{\mathbf{e}} = \vec{E}/|\vec{E}|$ associated with the electric field and create an unwanted phase difference between its components. However, in the regime of small losses, the configuration of the real part of the field still yields a well-defined skyrmion lattice, while being sufficiently larger than the imaginary part such that the latter is negligible [see full derivation in (*37*), section S.3]. In this regime, the skyrmion number S deviates slightly from its quantized value at zero losses owing to the weak breaking of the lattice translation symmetry, increasingly more so for unit cells that are farther away from the point of origin. Figure 3 gives a quantitative connection between the amount of loss and the robustness of the skyrmion number, showing the number of unit cells exhibiting $S > 0.99$. For example, a configuration in which 49 sites exhibit $S > 0.99$ is obtained if the electromagnetic waves creating it persist for roughly 400 periods before decaying—an easily achieved goal in many photonic systems.

A physical system capable of fulfilling the abovementioned requirements consists of surface plasmon polaritons (SPPs) (*38*)—electromagnetic surface waves existing at the interface between metallic and dielectric materials. SPPs only exist in TM polarization, and the phase difference between SPPs propagating along different directions can be easily controlled (*39*). Furthermore, they are, by design, an imperfect system, owing to the ohmic losses generated by the metal.

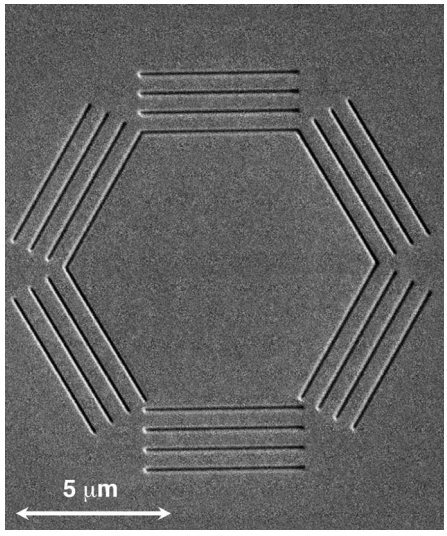


Fig. 4. Scanning electron microscopy image of the sample capable of generating an optical skyrmion lattice. The sample consists of six gratings creating a hexagon on a 200-nm-thick gold layer, deposited atop a 1-mm-thick glass substrate. The grating periodicity corresponds to the plasmonic wavelength (636 nm), and the bottom grating is displaced by half the plasmonic wavelength, to enable the necessary phase relations between the interfering waves. The optical skyrmion lattice is created at the center of the slit, through the SPPs propagating on the surface of the sample. For more information on the sample, see the materials and methods section in (*37*).

In our system (see Fig. 4), SPPs are excited at the interface of air and gold, resulting in both a small propagation decay and a transverse wave vector just slightly larger than the free-space wave number [$k_\parallel = (1.038 + \mathrm{i}0.003)k_0$]. Circularly polarized light impinges on a specially designed, hexagonally shaped coupling slit, exciting SPPs from each slit edge toward its center. The slit provides the same phase to the SPPs created by all edges, yet not exactly the same amplitude (owing to a different propagation length of the SPPs generated by two of the edges and their propagation loss). This results in a distortion of the skyrmion lattice, which, together with the finite structure, helps in examining its robustness.

In the experimental setup used to detect optical skyrmions [see (*37*) and fig. S1], a scattering near-field scanning optical microscope (Neaspec neaSNOM) enables phase-resolved measurement of the electric field normal to the surface by means of pseudoheterodyne interferometric detection (*40*). The ability to detect phase information is crucial, as it not only provides the full axial electric field but also allows us to perform a spatial Fourier transform (and its inverse) to filter out noises. The phase information, filtering ability, and high spatial resolution of the measurement enable the correct extraction of the transverse field components and, thus, of the skyrmion number.

The plasmonic wave number and applied boundary conditions create a plasmonic skyrmion lattice in the central part of the sample with characteristics typical of bubbles (Fig. 5). These characteristics are a result of the relatively small transverse wave vector, leading to an axial field component five times larger than the transverse ones. Although the real electric field is slightly

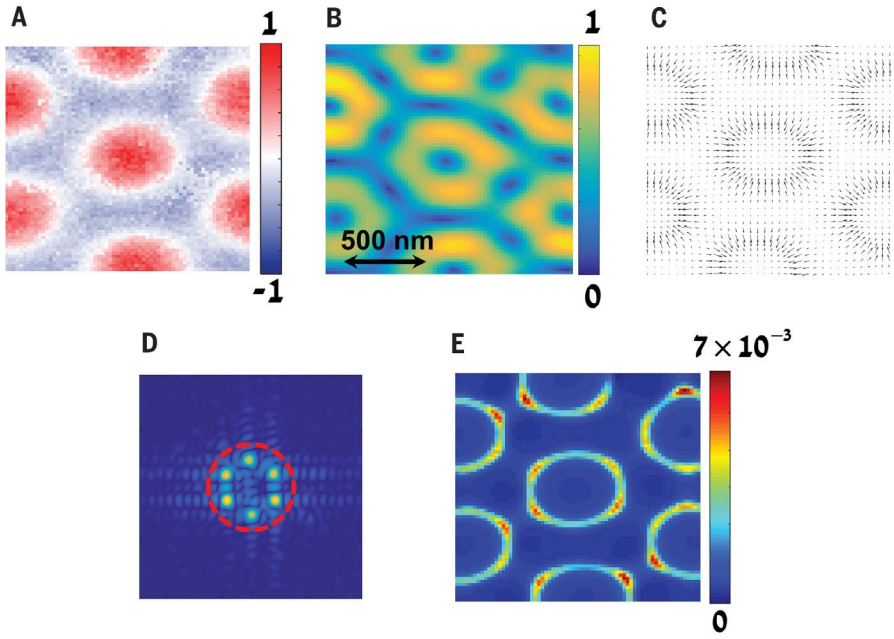


Fig. 5. Measurement of an optical skyrmion lattice created by SPPs. (**A**) Real part of the axial (out-of-plane) electric field at the center of the sample described in Fig. 3 (without noise filtering). (**B**) Amplitude of the transverse (in-plane) electric field. (**C**) Vector representation of the transverse electric field. (**D**) Amplitude of the Fourier decomposition of the longitudinal electric field. (**E**) Skyrmion number density map of the lattice. Scale bar is shown in (B). The complex value of the axial electric field was measured with 20-nm resolution by the system described in Fig. 3, and measurement noise was filtered with the low-pass filter, appearing in (D) as a red dashed circle. (B), (C), and (E) were then extracted from the data. The measured skyrmion number of a single lattice site was $S = 0.997 \pm 0.058$, which, together with the figures, shows a good match to the theoretical prediction. Actual number of lattice sites in our sample was 37. The color scales in (A) and (B) and the arrows in (C) are as defined for Fig. 1, A to C. The color scale in (E) represents the skyrmion number density *s*.

distorted, as expected (Fig. 5, A to C), it still shows similarity to the electric field configuration presented in Fig. 1. The extracted skyrmion number density at the center of the lattice (Fig. 5E) resembles that of the bubble-type skyrmion lattice shown in the inset of Fig. 2 (point i). Calculating the skyrmion number in each lattice site, we reach the result $S = 0.997 \pm 0.058$, thus demonstrating the robustness of the optical skyrmions.

Although not demonstrated experimentally in this work, the generation of a single skyrmion in the electric field of evanescent electromagnetic waves is possible as well [see (*37*), section S.5]. A single optical skyrmion may be achieved by engineering finite boundary conditions while still preserving a closed manifold, in a similar manner to schemes already suggested and observed in magnetic skyrmions (*41*, *42*).

Discussion

We show that a skyrmion lattice can be obtained as a solution to a linear plasmonic system by proper engineering of the boundary conditions and the transverse momentum of the electromagnetic waves. The topological invariant classifying the skyrmions in the lattice is protected by the lattice symmetry, and the lattice itself may be realized in any photonic system with evanescent waves—for example, planar waveguide modes and waves undergoing total internal reflection.

The generation of an optical skyrmion lattice paves the way toward inducing skyrmion lattices "on demand" in matter systems (e.g., cold atoms or dielectric particles in a fluid) through light-matter interactions. Namely, this will allow stimulated creation of skyrmions in matter, as opposed to skyrmions in all other systems [including optically excited skyrmions in chiral magnets (*43*)], which are spontaneously created.

Furthermore, optical nonlinearity in systems supporting evanescent waves (e.g., SPPs in gold or nonlinear dielectric waveguides) could give rise to soliton-like skyrmion states exhibiting a topologically protected skyrmion number, which will be robust against external perturbation.

Here we focused on TM-polarized electromagnetic waves, which exhibited a skyrmion lattice in the electric field. Arranging transverse-electric polarized waves in a similar way, by using a dielectric waveguide or by total internal reflection, would create skyrmions in the magnetic field of the electromagnetic wave, with the potential to stimulate a skyrmion lattice not only electrically but also magnetically by using either femtosecond pulsed excitations or submillikelvin temperatures.

Optical skyrmion lattices could also bring about new optical effects, highlighting their potential applications in optical information processing, storage, and transfer. For example, in light-emission processes such as fluorescence or harmonic generation, the specific field configuration of optical skyrmions would result in the simultaneous emission of all possible polarizations in a structured, phase-locked, and coherent way. Alternatively, probe beams could nonlinearly interact with the skyrmion lattice via Kerr nonlinearity and be affected by a complex, polarization-dependent Berry curvature.

REFERENCES AND NOTES

1. J. Bardeen, M. J. Stephen, *Phys. Rev.* **140**, A1197–A1207 (1965).
2. M. R. Matthews *et al.*, *Phys. Rev. Lett.* **83**, 2498–2501 (1999).
3. T. W. B. Kibble, *J. Phys. Math. Gen.* **9**, 1387–1398 (1976).
4. L. Allen *et al.*, *Phys. Rev. A* **45**, 8185–8189 (1992).
5. G. Blatter *et al.*, *Rev. Mod. Phys.* **66**, 1125–1388 (1994).
6. J. M. Kosterlitz, D. J. Thouless, *J. Phys. C Solid State Phys.* **6**, 1181–1203 (1973).
7. T. H. R. Skyrme, *Nucl. Phys.* **31**, 556–569 (1962).
8. U. Al Khawaja, H. Stoof, *Nature* **411**, 918–920 (2001).
9. J. Fukuda, S. Žumer, *Nat. Commun.* **2**, 246 (2011).
10. S. Mühlbauer *et al.*, *Science* **323**, 915–919 (2009).
11. X. Z. Yu *et al.*, *Nature* **465**, 901–904 (2010).
12. N. Romming *et al.*, *Science* **341**, 636–639 (2013).
13. D. Maccariello *et al.*, *Nat. Nanotechnol.* **13**, 233–237 (2018).
14. A. Fert, V. Cros, J. Sampaio, *Nat. Nanotechnol.* **8**, 152–156 (2013).
15. N. Nagaosa, Y. Tokura, *Nat. Nanotechnol.* **8**, 899–911 (2013).
16. X. Z. Yu *et al.*, *Nat. Commun.* **3**, 988 (2012).
17. J. Sampaio, V. Cros, S. Rohart, A. Thiaville, A. Fert, *Nat. Nanotechnol.* **8**, 839–844 (2013).
18. I. Kézsmárki *et al.*, *Nat. Mater.* **14**, 1116–1122 (2015).
19. C. Kooy, U. Enz, *Philips Res. Rep.* **15**, 7–29 (1960).
20. C. Moutafis *et al.*, *Phys. Rev. B* **79**, 224429 (2009).
21. Z. Wang *et al.*, *Nature* **461**, 772–775 (2009).
22. K. Fang, Z. Yu, S. Fan, *Nat. Photonics* **6**, 782–787 (2012).
23. A. B. Khanikaev *et al.*, *Nat. Mater.* **12**, 233–239 (2013).
24. M. C. Rechtsman *et al.*, *Nature* **496**, 196–200 (2013).
25. M. Hafezi, S. Mittal, J. Fan, A. Migdall, J. M. Taylor, *Nat. Photonics* **7**, 1001–1005 (2013).
26. B. Bahari *et al.*, *Science* **358**, 636–640 (2017).
27. G. Harari *et al.*, *Science* **359**, eaar4003 (2018).
28. M. A. Bandres *et al.*, *Science* **359**, eaar4005 (2018).
29. N. B. Simpson *et al.*, *Opt. Lett.* **22**, 52–54 (1997).
30. M. S. Soskin, V. Gorshkov, M. Vasnetsov, J. Malos, N. Heckenberg, *Phys. Rev. A* **56**, 4064–4075 (1997).
31. F. Flossmann *et al.*, *Phys. Rev. Lett.* **95**, 253901 (2005).
32. Y. Gorodetski *et al.*, *Phys. Rev. Lett.* **101**, 043903 (2008).
33. G. Spektor *et al.*, *Science* **355**, 1187–1191 (2017).
34. E. Ostrovsky *et al.*, *Optica* **5**, 283–288 (2018).
35. M. Gilles *et al.*, *Nat. Photonics* **11**, 102–107 (2017).
36. T. Van Mechelen, Z. Jacob, *Optica* **3**, 118–126 (2016).
37. Available in the materials and methods or supplementary text, as part of the supplementary materials.
38. S. A. Maier, *Plasmonics: Fundamentals and Applications* (Springer, 2007).
39. G. Spektor *et al.*, *Opt. Express* **23**, 32759–32765 (2015).
40. N. Ocelic *et al.*, *Appl. Phys. Lett.* **89**, 101124 (2006).
41. L. Sun *et al.*, *Phys. Rev. Lett.* **110**, 167201 (2013).
42. D. A. Gilbert *et al.*, *Nat. Commun.* **6**, 8462 (2015).
43. G. Berruto *et al.*, *Phys. Rev. Lett.* **120**, 117201 (2018).

ACKNOWLEDGMENTS

We thank S. Dolev for his help in fabrication of the measured samples. S.T. also acknowledges the generous support of the Jacobs Foundation and I. Khanonkin's help in performing simulations. **Funding:** This research was supported by "Circle of Light," Israeli Centers for Research Excellence (I-CORE), through the Israel Science Foundation (ISF), grant no. 1802/12; and by the European Research Council (Horizon 2020 program), grant no. 639172. **Author contributions:** S.T., B.G., and G.B. conceived the project; S.T. and K.C. patterned the samples; E.O., S.T., and K.C. performed the measurements; and S.T., N.L., and G.B. performed simulations and analytical calculations. All authors took part in preparing the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are available in the main text or the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6406/993/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S5
Reference (*44*)

29 April 2018; accepted 3 July 2018
Published online 19 July 2018
10.1126/science.aau0227

REPORT

COMBUSTION CHEMISTRY

Resonance-stabilized hydrocarbon-radical chain reactions may explain soot inception and growth

K. O. Johansson[1]*, M. P. Head-Gordon[2,3], P. E. Schrader[1], K. R. Wilson[3], H. A. Michelsen[1]*

Mystery surrounds the transition from gas-phase hydrocarbon precursors to terrestrial soot and interstellar dust, which are carbonaceous particles formed under similar conditions. Although polycyclic aromatic hydrocarbons (PAHs) are known precursors to high-temperature carbonaceous-particle formation, the molecular pathways that initiate particle formation are unknown. We present experimental and theoretical evidence for rapid molecular clustering–reaction pathways involving radicals with extended conjugation. These radicals react with other hydrocarbon species to form covalently bound complexes that promote further growth and clustering by regenerating resonance-stabilized radicals through low-barrier hydrogen-abstraction and hydrogen-ejection reactions. Such radical–chain reaction pathways may lead to covalently bound clusters of PAHs and other hydrocarbons that would otherwise be too small to condense at high temperatures, thus providing the key mechanistic steps for rapid particle formation and surface growth by hydrocarbon chemisorption.

Soot is produced during incomplete combustion of hydrocarbon fuels. Carbonaceous interstellar particles, containing ~70% of the carbon in interstellar space, are formed under similar chemical conditions (*1*–*5*). Terrestrial soot has enormous impact on human health and the environment (*6*); it plays a major role in deaths attributed to air pollution worldwide (*7*) and is an important contributor to global warming (*8*). Notably, the identity of the chemical mechanisms for soot and interstellar dust formation is a scientific puzzle analogous to a former long-standing challenge of understanding particle nucleation in Earth's atmosphere. The key pathway to new-particle formation and growth from biogenic vapors was only recently discovered (*9*, *10*).

There is a large body of evidence linking polycyclic aromatic hydrocarbon (PAH) species to soot formation (*11*–*13*). Key missing steps in the understanding of soot formation include inception (i.e., the production of hydrocarbon clusters that mark a transition to the condensed phase) and the continuation of rapid growth by hydrocarbon addition to these incipient particles (i.e., hydrocarbon clusters). There are two main classes of mechanisms for soot inception under active research: physical bonding of PAHs by dispersive or van der Waals forces into stacked clusters and formation of covalently bound clusters (*12*). At temperatures where soot inception occurs (≥1450 ± 250 K), the PAHs typically present in appreciable quantity contain two to five aromatic rings and are far too volatile to nucleate by van der Waals forces (*12*); a long history of research indicates a need for a chemical mechanism that rapidly bonds PAHs covalently at these temperatures. No detailed chemical pathways have been posited, however, and current chemical models for covalent-cluster formation generally include proxy reactions to quickly and irreversibly bond highly stable PAHs in violation of the second law of thermodynamics (*12*). Such reactions tend to necessitate repeated activation of stable PAHs through H abstraction via substantial barriers (forcing rates to be too slow), yield nearly uniform growth among all PAHs instead of cluster formation, and quench the radical pool (*12*). The proposed mechanisms and associated issues are similar for particle-surface growth. These issues are also relevant to the understanding of stardust formation near carbon-rich stars because the chemistry and precursor species of interstellar-dust formation are thought to be very similar to those of terrestrial soot (*1*–*4*).

Here we present theoretical and experimental evidence for a radical-driven hydrocarbon-clustering mechanism that could provide a physically viable route to particle formation and growth. This mechanism involves chain reactions of resonance-stabilized radical (RSR) species that can form covalent bonds with a wide variety of hydrocarbons to produce a clustered product with RSR character. The mechanism is initiated and propagated by RSRs that require an unpaired electron for extended conjugation. Products of reactions of these RSRs with other radical and closed-shell hydrocarbons readily regenerate RSRs that maintain extended conjugation via an unpaired electron. Hence, loss and gain of extended conjugation promote these RSR-driven radical-chain reactions. This radical-regeneration process maintains a pool of radicals that can undergo molecular growth and act as clustering nuclei for particle formation. We refer to this mechanism as clustering of hydrocarbons by radical-chain reactions (CHRCR). Because larger carbonaceous particles are likely to have surface-radical sites with chemical traits similar to those of the RSRs, we hypothesize that these pathways may also contribute to particle-surface growth under high-temperature conditions.

Figure 1 provides an overview of the CHRCR mechanism in three stages. The first stage (Fig. 1A) increases the size of an RSR through acetylene (C_2H_2) or vinyl (C_2H_3) addition, via radical-chain reactions. We chose to start this stage with cyclopentadienyl, but propargyl (C_3H_3 at 39 u) is also an RSR, which generates cyclopentadienyl through reaction with acetylene (*14*), and thus precedes cyclopentadienyl in the sequence. The second stage (Fig. 1B) involves hydrocarbon clustering via radical-chain reactions initiated by the RSRs from the first stage. The third stage (Fig. 1C) is an extension of the second and might promote particle surface growth by RSR-driven radical-chain reactions at sites on the particle surface.

The first stage (Fig. 1A) involves molecular growth of the pool of RSRs that act as seeds for initiating hydrocarbon clustering. Our electronic-structure calculations predict a sequence of RSRs readily generated through radical-chain reactions. Each reaction adds acetylene (C_2H_2) or vinyl (C_2H_3) to an RSR and generates a new RSR (see supplementary materials). Hence, these radicals are not consumed when reacting with C_2H_2 and C_2H_3. Instead they grow to larger species with similar chemical traits and could, therefore, survive long enough to act as clustering seeds for other hydrocarbons. The species generated by these reactions, starting from 65 u, have masses of 91, 115, 141, 165, 189, 215, 239, and 263 u and, likely, higher values. These masses are frequently observed experimentally in condensed samples extracted from flames (*13*, *15*–*18*). Acetylene is generally the most abundant hydrocarbon under high-temperature conditions in space and on Earth (*3*, *11*). This stage of the mechanism thus has strong theoretical and experimental support.

We used a vacuum ultraviolet aerosol mass spectrometer (VUV-AMS) to measure these species on particles extracted from flames. A mass spectrum highlighting the RSR sequence is shown in Fig. 2. The sample was extracted from an atmospheric-pressure laminar premixed ethylene-oxygen flame. Laminar premixed flames provide stable and reproducible flame conditions with controllable fuel to oxidizer ratios. Fuel-rich premixed combustion is relevant to the first stages of soot

[1]Combustion Research Facility, Sandia National Laboratories, Livermore, CA 94550, USA. [2]Department of Chemistry, University of California Berkeley, Berkeley, CA 94720, USA. [3]Chemical Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.
*Corresponding author. Email: okjohans@gmail.com (K.O.J.); hamiche@sandia.gov or hamichelsen2@gmail.com (H.A.M.)

formation under diesel-engine conditions (*19*), and ethylene is a standard fuel for soot-formation studies because of its simple structure and high propensity to produce soot. Most natural flames are formed under diffusion-controlled conditions. In the supplementary materials, we present similar results from both diffusion and premixed flames with several different fuels. Concentrations suggested by peak intensities for the RSRs are biased low relative to heavier closed-shell species because many radical-quenching reactions preclude detection or yield signal at different masses.

Whereas prior work has assumed that the mass peak at 91 u (R91) stems from benzyl, we identified R91 as vinylcyclopentadienyl from its ionization energy in both ethylene- and acetylene-fueled flames (see supplementary materials). Likely formation pathways include vinyl or acetylene addition to cyclopentadienyl (Fig. 1A). The large numbers of isomers for species heavier than R91 preclude experimental identification; however, the mass sequence highlighted in Fig. 2 is consistent with the predicted RSRs.

The second stage of the mechanism (Fig. 1B) describes clustering of available hydrocarbons by the initiator radicals produced in the first stage. Figure 3 shows an example of a pathway commencing with reaction between indenyl and phenyl to form a covalently bound closed-shell dimer (i.e., σ-dimer). Indenyl ($C_9H_7^{\bullet}$) is a resonantly stabilized π-radical whose extended conjugation depends on maintaining an unpaired π electron. Hence, indenyl undergoes partial loss of conjugation during σ-dimerization (the site of phenyl attack changes hybridization from sp^2 to sp^3). The dimerization process is barrierless, but because indenyl loses conjugation during dimerization, the energy is lowered substantially less than a typical C-C bond energy. The loss of conjugation means that the barrier to subsequent H abstraction from the α carbon of the five-membered ring, i.e., the carbon atom to which phenyl attached (using R=H), can be as low as ~4.5 kcal/mol (Fig. 3 shows two abstraction sites) because H loss leads to extended conjugation. This energy landscape also promotes direct H ejection from the chemically activated

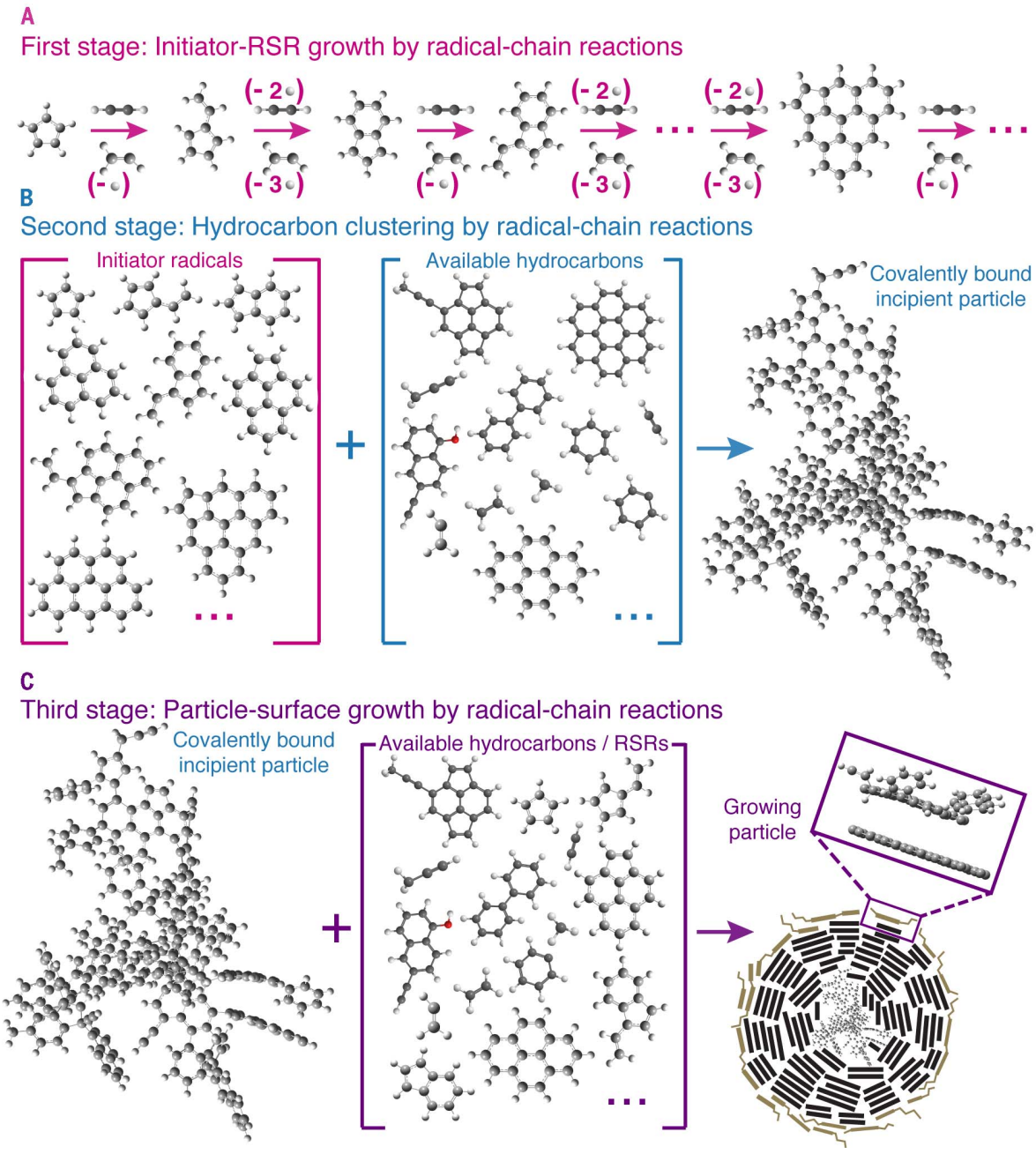


Fig. 1. Schematic overview of the clustering of hydrocarbons by radical-chain reactions (CHRCR) mechanism. (**A**) The CHRCR mechanism is initiated and propagated by resonantly stabilized radicals sequentially generated through radical-chain reactions involving acetylene or vinyl. (**B**) To form an incipient particle, these RSRs can cluster a wide range of hydrocarbons, including radicals, stable PAHs, and unsaturated aliphatic species, through radical-chain reactions fueled by loss and gain of extended conjugation. (**C**) Cyclopentadienyl-type moieties on cluster surfaces are posited to further propagate growth via the CHRCR mechanism.

dimer. H-atom abstraction or ejection generates a new RSR, similar in character to indenyl, whose extended conjugation stabilizes its unpaired electron and reduces the bond length between the dimerized entities in Fig. 3 from ~1.51 to ~1.45 Å. Several of the CHRCR reaction pathways could proceed without radical consumption or with net radical production (see supplementary materials).

The barrierless dimerization process and low H-abstraction barrier provide the energy landscape required for fast reactions. Dimerization and H abstraction together release more than 100 kcal/mol (zero-point corrected electronic energies), minimizing reversibility. In addition to reacting with a PAH radical, indenyl can attack an unsaturated aliphatic species (see supplementary materials) or a nonradical aromatic ring. The latter reaction, however, leads to loss of conjugation in both indenyl and the aromatic ring, yielding a weakly bound dimer that requires fast stabilization via H loss.

These reactions represent a repeatable class of pathways for cluster formation. The clustering is accelerated by adding RSRs whose extended conjugation requires an unpaired electron and benefits from previously proposed gas-phase growth mechanisms, which can expand the sizes of clustered species before and after clustering. Each clustered species adds edge sites that can become activated through H abstraction. This scheme concentrates clustering on a limited number of nucleation seeds and readily attaches species too volatile to condense via dispersive forces at high temperatures.

The reactants for this second stage of the mechanism are commonly observed under flame conditions. In addition, many experimental studies have demonstrated the formation of small (1 to 6 nm) incipient particles with carbon-to-hydrogen ratios of ~2 composed of randomly ordered aromatic structures with some aliphatic character (*20*), which are consistent with the clusters predicted by the CHRCR mechanism. This stage of the mechanism is thus supported by theoretical and experimental results. In addition, the CHRCR mechanism clusters both aromatic and aliphatic species and may explain the aliphatic and aromatic content observed spectroscopically in

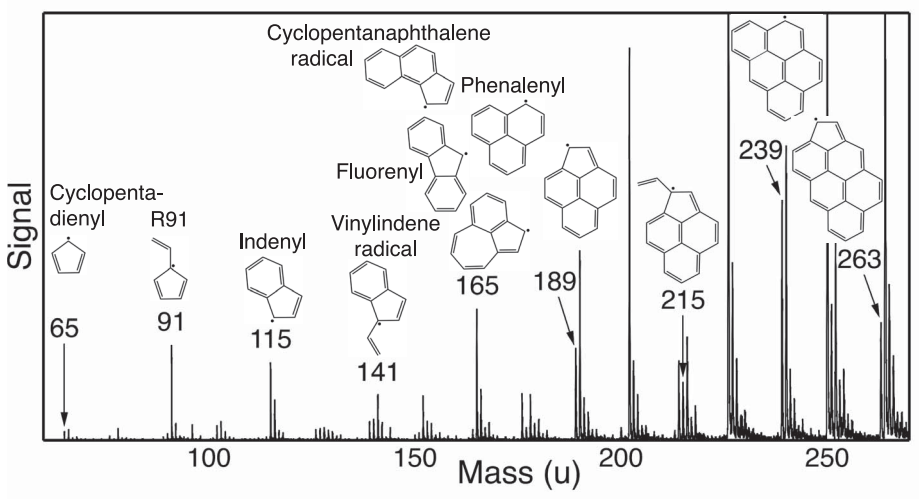


Fig. 2. VUV-AMS spectrum demonstrating a sequence of radicals. This VUV-AMS spectrum was recorded on a sample extracted from an atmospheric-pressure laminar premixed ethylene-oxygen flame using a photon energy of 9.4 eV. Masses are indicated for species posited to drive the CHRCR mechanism. Our best estimates of the predominant structures are shown. Four hypothesized isomers are shown for mass 165 u. There are many more potential isomers for higher radical masses not shown.

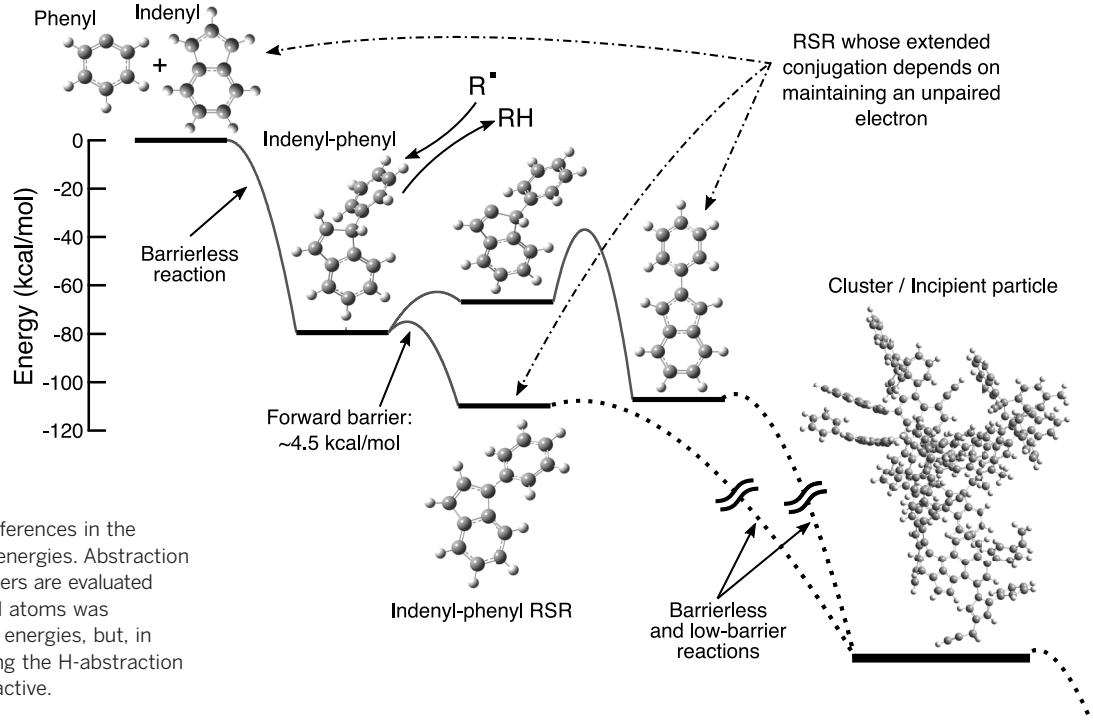


Fig. 3. Representative potential energy surface for hydrocarbon clustering. The potential energy diagram shows formation of an indenyl-phenyl σ-dimer and subsequent H abstraction, which stabilizes the bond and regenerates an RSR whose extended conjugation requires an unpaired electron (see supplementary materials for details on the calculations). These reactions are examples of possible steps in the hydrocarbon-clustering scheme proposed. Phenyl represents an arbitrary PAH radical; the cluster is shown for illustrative purposes. The *y* axis displays differences in the zero-point–corrected electronic energies. Abstraction steps, energy changes, and barriers are evaluated with R=H. The total number of H atoms was identical for all of the calculated energies, but, in the figure, species involved during the H-abstraction reactions are only shown when active.

carbonaceous particles formed in interstellar space and in laboratory experiments that mimic the interstellar and circumstellar conditions where stardust forms (*1*, *2*, *4*, *5*).

Soot particles that have grown and chemically evolved at high temperature are composed of aggregates of quasi-spherical primary particles with diameters of 10 to 50 nm. These primary particles often have disordered carbon cores surrounded by more ordered shells of graphite crystallites (*21*, *22*). The particle core is 1 to 4 nm in diameter (*21*, *22*), which is consistent with the size range of clusters likely formed by the CHRCR mechanism, indicating that such clusters could seed further particle growth.

The third stage of the mechanism involves growth beyond the initial cluster (Fig. 1C). We hypothesize that, once clusters have formed, RSR moieties at particle surfaces could provide sites for chemisorption of small hydrocarbons that are abundant but too volatile to condense on the surface via dispersive forces. Thus, CHRCR pathways could potentially explain recent results showing that the surfaces of soot particles may have less aromatic and more aliphatic character than the rest of the graphitic shell (*12*, *23*).

The CHRCR mechanism overcomes the major problem of insufficient abundances of viable PAHs because it can account for clustering of a wide range of hydrocarbon sizes and structures. This mechanism thus relies on abundant small hydrocarbons observed under conditions where carbonaceous particles are formed. Because the CHRCR mechanism regenerates radicals and is likely to eject free H atoms as it proceeds, it concentrates clustering on certain species and minimizes radical quenching, which slows other mechanisms involving radical reactions.

The CHRCR mechanism is relevant to rich hydrocarbon combustion where soot is formed, the outflow of carbon-rich stars where carbonaceous dust is formed, and hydrocarbon-pyrolysis conditions; it is generally applicable for conditions under which hydrocarbon radicals can form and react by high-temperature or photolytic processes. Future studies are needed to investigate reaction rates and yields of these RSRs, rates and identities of nucleated particles, and rates and mechanisms of particle-surface growth under relevant conditions to verify these results.

REFERENCES AND NOTES

1. A. Léger, J. L. Puget, *Astron. Astrophys.* **137**, L5–L8 (1984).
2. L. J. Allamandola, A. G. G. M. Tielens, J. R. Barker, *Astrophys. J.* **290**, L25–L28 (1985).
3. A. G. G. M. Tielens, in *Carbon in the Galaxy*, J. C. Tarter, S. Chang, D. DeFrees, Eds. (NASA Conference Publication, 1990), vol. 3061, pp. 59–111.
4. A. G. G. M. Tielens, *Annu. Rev. Astron. Astrophys.* **46**, 289–337 (2008).
5. B. Günay *et al.*, *Mon. Not. R. Astron. Soc.* **479**, 4336–4344 (2018).
6. B. J. Finlayson-Pitts, J. N. Pitts Jr., *Science* **276**, 1045–1052 (1997).
7. WHO, (World Health Organization, www.who.int/mediacentre/news/releases/2014/air-pollution/en/, 2014).
8. T. C. Bond *et al.*, *J. Geophys. Res. Atmos.* **118**, 5380–5552 (2013).
9. F. Bianchi *et al.*, *Science* **352**, 1109–1112 (2016).
10. M. Ehn *et al.*, *Nature* **506**, 476–479 (2014).
11. M. Frenklach, *Phys. Chem. Chem. Phys.* **4**, 2028–2037 (2002).
12. H. Wang, *Proc. Combust. Inst.* **33**, 41–67 (2011).
13. K. H. Homann, *Angew. Chem. Int. Ed.* **37**, 2434–2451 (1998).
14. V. D. Knyazev, I. R. Slagle, *J. Phys. Chem. A* **106**, 5613–5617 (2002).
15. B. Öktem, M. P. Tolocka, B. Zhao, H. Wang, M. V. Johnston, *Combust. Flame* **142**, 364–373 (2005).
16. S. A. Skeen *et al.*, *J. Aerosol Sci.* **58**, 86–102 (2013).
17. K. O. Johansson *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 8374–8379 (2016).
18. R. A. Dobbins, R. A. Fletcher, H.-C. Chang, *Combust. Flame* **115**, 285–298 (1998).
19. M. P. B. Musculus, P. C. Miles, L. M. Pickett, *Pror. Energy Combust. Sci.* **39**, 246–283 (2013).
20. H. A. Michelsen, *Proc. Combust. Inst.* **36**, 717–735 (2017).
21. T. Ishiguro, Y. Takatori, K. Akihama, *Combust. Flame* **108**, 231–234 (1997).
22. P. Toth *et al.*, *Combust. Flame* **162**, 2422–2430 (2015).
23. K. O. Johansson, F. El Gabaly, P. E. Schrader, M. F. Campbell, H. A. Michelsen, *Aerosol Sci. Technol.* **51**, 1333–1344 (2017).

ACKNOWLEDGMENTS

We thank B. Rude and D. Taube for assistance at the ALS and R. Bambha for help improving the paper's clarity. **Funding:** This work was funded by the U.S. Department of Energy (DOE), Office of Science, Office of Basic Energy Sciences (BES), Chemical Sciences, Geosciences, and Biosciences Division, Gas Phase Chemical Physics Program. M.P.H.-G. and K.R.W. are supported by this program under Contract no. DE-AC02-05CH11231. The ALS at LBNL is supported by the Director, DOE BES, under Contract no. DE-AC02-05CH11231. SNL is a multimission laboratory, managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the DOE's National Nuclear Security Administration under Contract DE-NA0003525. The views expressed in this article do not necessarily represent those of the DOE or U.S. Government. **Author contributions:** K.O.J., H.A.M., M.P.H.-G., and K.R.W. wrote the paper. K.O.J. and H.A.M. conceived of the project and wrote the original draft of the paper. K.O.J., P.E.S., K.R.W., and H.A.M. performed the experimental work. K.O.J. analyzed the mass spectra and performed the calculations. M.P.H.-G. helped guide the calculations. K.R.W. built the aerosol mass spectrometer. H.A.M. managed the project. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The mass spectra shown in Fig. 2 and figs. S1 to S3 are provided in a separate file (Data File S1).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6406/997/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S11
References (*24*–*36*)
Data File S1

15 February 2018; accepted 13 July 2018
10.1126/science.aat3417

PLASMA ASTROPHYSICS

Direct measurements of two-way wave-particle energy transfer in a collisionless space plasma

N. Kitamura[1,2]*, M. Kitahara[3], M. Shoji[4], Y. Miyoshi[4], H. Hasegawa[1], S. Nakamura[5], Y. Katoh[3], Y. Saito[1], S. Yokota[6], D. J. Gershman[7], A. F. Vinas[7,8], B. L. Giles[7], T. E. Moore[7], W. R. Paterson[7], C. J. Pollock[9], C. T. Russell[10], R. J. Strangeway[10], S. A. Fuselier[11,12], J. L. Burch[11]

Particle acceleration by plasma waves and spontaneous wave generation are fundamental energy and momentum exchange processes in collisionless plasmas. Such wave-particle interactions occur ubiquitously in space. We present ultrafast measurements in Earth's magnetosphere by the Magnetospheric Multiscale spacecraft that enabled quantitative evaluation of energy transfer in interactions associated with electromagnetic ion cyclotron waves. The observed ion distributions are not symmetric around the magnetic field direction but are in phase with the plasma wave fields. The wave-ion phase relations demonstrate that a cyclotron resonance transferred energy from hot protons to waves, which in turn nonresonantly accelerated cold He^+ to energies up to ~2 kilo–electron volts. These observations provide direct quantitative evidence for collisionless energy transfer in plasmas between distinct particle populations via wave-particle interactions.

Wave-particle interactions are thought to play a crucial role in energy transfer in collisionless space plasmas in which the motion of charged particles is controlled by electromagnetic fields. In Earth's magnetosphere, electrons with energies on the order of a few to several tens of kilo–electron volts spontaneously generate electromagnetic electron cyclotron waves, called chorus emissions. Cyclotron resonant interaction with such waves and the resulting acceleration of electrons with energies on the order of several hundred kilo–electron volts are a leading candidate for the generation of relativistic electrons (on the order of mega–electron volts), which constitute the Van Allen radiation belt (*1*–*3*). Electromagnetic waves near the ion cyclotron frequency can accelerate ions through cyclotron resonance in the polar region (*4*), leading to the loss of O^+ from Earth's atmosphere. Electromagnetic ion cyclotron (EMIC) waves, generated spontaneously by hot ions in the equatorial magnetosphere, can cause loss of energetic ions via cyclotron resonant scattering, contributing to decay of geomagnetic storms (*5*). These waves can also induce quick loss of "satellite-killer" mega–electron volt electrons in the radiation belts during geomagnetic storms, limiting the threat that they pose to satellites (*6*, *7*). A quantitative understanding of wave-particle interactions and energy transfer between particle populations would therefore inform various space plasma phenomena such as the radiation belt, geomagnetic storms, auroral particle precipitation, and atmospheric loss from planets.

The coexistence of waves and accelerated particles (or particle populations that have free energy for wave growth) has been studied for decades in the magnetosphere (*8*–*10*). However, such coexistence does not necessarily indicate that energy is transferred between them at the observation site and time. In most situations, moving particles interact gradually with propagating waves in a spatially extended region, and it is not realistic to track a certain particle or wave packet with spacecraft. Thus, detecting local energy transfer between the fields and particles is necessary to quantitatively evaluate the magnitude of any interaction. Flux modulation of auroral precipitating electrons that may be related to cyclotron interactions with electrostatic waves was detected in the ionosphere (*11*). For direct quantitative measurements of the energy exchange between particle and electromagnetic waves via cyclotron interactions, the Wave-Particle Interaction Analyzer method that uses observed waveform and nonuniformity of particles around the magnetic field lines was proposed (*12*, *13*). Using this method, the detection of energy transfer from ions to waves via nonlinear cyclotron interactions has been achieved recently with in situ measurements with a temporal resolution of ~40 wave periods (*14*). However, the limited field of view and temporal resolution of their ion detectors did not allow observation of details of the interaction during the course of wave evolution (growth or decay), as characterized by temporal variations of the wave amplitude. We present direct evidence of energy transfer between two distinct particle populations through two concurrent cyclotron interactions based on quantitative measurements of the interactions, with a temporal resolution as high as one wave period.

The four Magnetospheric Multiscale (MMS) spacecraft (*15*) observed EMIC waves around 12:20 universal time (UT) on 1 September 2015 in the dusk-side magnetosphere (Fig. 1A and fig. S1). Because the spacecraft separation was smaller than both the wavelength estimated from the dispersion relation (fig. S2) and the cyclotron radius of hot H^+ (*16*), we use data averaged over all four spacecraft unless otherwise noted. We used the background magnetic field ($\mathbf{B}_0$) (<0.05 Hz) to define the magnetic field–aligned coordinates. The wave component of the magnetic field in the frequency range from 0.05 to 0.15 Hz, which is around the peak of the wave power (fig. S3), was derived as the wave magnetic field ($\mathbf{B}_{wave}$) (Fig. 2A). The perpendicular component of the wave electric field ($\mathbf{E}_{wave}$) in the same frequency range (Fig. 2B) was derived from cold ion motion (*16*). The field-aligned component of the Poynting flux [$\mathbf{S} = (\mathbf{E}_{wave} \times \mathbf{B}_{wave})/\mu_0$, where μ_0 is the vacuum permeability] was negative for most of the time interval, so the wave was propagating antiparallel to $\mathbf{B}_0$ (Figs. 1B and 2C).

The energy transfer rate via cyclotron-type interactions between cyclotron waves and ions was calculated as the dot product of $\mathbf{E}_{wave}$ and the ion current ($\mathbf{j}_i$) perpendicular to $\mathbf{B}_0$. In the case of resonant interactions between ions and waves, the current is called the resonant current (*17*). Over several energy and pitch-angle ranges, $\mathbf{j}_i$ was calculated by using burst data from the Fast Plasma Investigation Dual Ion Spectrometer (FPI-DIS) on MMS (*18*) with a time resolution of 150 ms, which is ~1/100 of the wave period (*16*). In a magnetized plasma, we expect the particles to be uniformly distributed around the magnetic field lines and call this uniformity "gyrotropy." The measured nonuniformity, or agyrotropy of ions, corresponding to $\mathbf{j}_i \neq 0$ causes an imbalance between the ions accelerated and decelerated by $\mathbf{E}_{wave}$, if $\mathbf{j}_i \cdot \mathbf{E}_{wave} \neq 0$. Thus, the agyrotropy and $\mathbf{j}_i \cdot \mathbf{E}_{wave}$ determine the net energy transfer for each part of the energy and pitch-angle ranges. Although this concept is the same as the Wave-Particle Interaction Analyzer method developed and used in previous works (*12*–*14*), the full-sky field of view of FPI-DIS enabled fast measurements of instantaneous $\mathbf{j}_i$ and thus of energy transfer rate ($\mathbf{j}_i \cdot \mathbf{E}_{wave}$).

First-order cyclotron resonance occurs when the resonance condition $V_{R\_i} = (\omega - \Omega_i)/k_{para}$ is satisfied, where V_{R_i} is the resonance velocity for ions, ω is the angular frequency of the left-hand polarized cyclotron wave, Ω_i is the ion

[1]Institute of Space and Astronautical Science, Japan Aerospace Exploration Agency, Sagamihara, Japan. [2]Department of Earth and Planetary Science, Graduate School of Science, the University of Tokyo, Tokyo, Japan. [3]Department of Geophysics, Graduate School of Science, Tohoku University, Sendai, Japan. [4]Institute for Space-Earth Environmental Research (ISEE), Nagoya University, Nagoya, Japan. [5]Research Institute for Sustainable Humanosphere (RISH), Kyoto University, Uji, Japan. [6]Department of Earth and Space Science, Graduate School of Science, Osaka University, Toyonaka, Japan. [7]NASA Goddard Space Flight Center, Greenbelt, MD, USA. [8]Department of Physics, American University, Washington, DC, USA. [9]Denali Scientific, Fairbanks, AK, USA. [10]Institute of Geophysics and Planetary Physics, University of California, Los Angeles, CA, USA. [11]Southwest Research Institute, San Antonio, TX, USA. [12]University of Texas at San Antonio, San Antonio, TX, USA.
*Corresponding author. Email: kitamura@eps.s.u-tokyo.ac.jp

cyclotron frequency, and k_{para} is the wave number parallel to $\mathbf{B}_0$. The subscripts i indicate H^+ or He^+ ions. This condition is met when the angular frequency of $\mathbf{E}_{wave}$ and $\mathbf{B}_{wave}$ seen by ions with a parallel velocity V_{R_i} becomes equal to Ω_i because of the Doppler shift. In the MMS observation considered here, because the wave was propagating antiparallel to $\mathbf{B}_0$ and $\omega < \Omega_{He+}$, the resonance condition can be satisfied for H^+ or He^+ with pitch angles smaller than 90°.

Around 12:18:30 UT, 15-s averages of $\mathbf{j}_i \cdot \mathbf{E}_{wave}$ reached −0.3 pW m^{-3} for ions with energies 14 to 30 keV and pitch angles 33.25° to 78.75° (Fig. 2D), where the resonance condition for H^+ was satisfied. We confirmed that H^+ is the dominant ion species in this energy range (fig. S3) (*16*). For ions with pitch angles 101.25° to 146.75°, which did not satisfy the resonance condition, the averaged $\mathbf{j}_i \cdot \mathbf{E}_{wave}$ stayed much closer to zero (Fig. 2D), even though the pitch-angle distributions of the ions were almost symmetric about 90° (Fig. 2E). These results demonstrate that the energy of H^+ was being transferred to the cyclotron wave by the cyclotron resonance. These features were consistently observed by each of the four spacecraft, attesting to the robustness of the results (fig. S4).

A gyro phase versus time plot of differential energy fluxes is shown in Fig. 2F for ions with energies 14 to 30 keV and pitch angles 33.25° to 78.75°. To emphasize the agyrotropy, we normalized the values using the gyro-averaged values at each time (Fig. 2G). Two types of agyrotropy were seen. The first is stable in gyro angle (~12:17:20 to 12:18:10 and ~12:21:00 to 12:22:10 UT) and is related to the spatial gradient of ion fluxes, and the second is rotating (~12:18:10 to 12:19:15 UT). In the former case, $\mathbf{j}_i \cdot \mathbf{E}_{wave}$ cancels out over one complete wave period, and so the agyrotropy does not contribute to the net energy transfer. The latter case was investigated in more detail, by sorting the data using the relative phase angle (ζ), which is the gyro phase relative to the rotating $\mathbf{B}_{wave}$ (fig. S5). The resulting ζ versus time plot is shown in Fig. 2H. Relatively low ion fluxes were detected near the direction parallel to, and relatively high ion fluxes were detected near the direction antiparallel to, $\mathbf{E}_{wave}$, which remained at ζ ~ 90° around 12:18:10 to 12:18:45 UT. This agyrotropy rotating with $\mathbf{B}_{wave}$ and $\mathbf{E}_{wave}$ leads to the negative $\mathbf{j}_i \cdot \mathbf{E}_{wave}$ (Fig. 2D). Above 14 keV, a significant dip (~20 to 50% decrease from peak flux) at ζ ~ 90° can be seen in multiple pitch-angle bins, each consisting of individual measurements, in the ζ distribution of the differential energy fluxes (Fig. 3, A to D). As a quantitative measure of energy transfer, we computed gyro phase-averaged energy gain per H^+ ion perpendicular to $\mathbf{B}_0$ for each bin by dividing $\mathbf{j}_i \cdot \mathbf{E}_{wave}$ by the partial number density (Fig. 3E). Energy loss rates (negative energy gain) of up to ~80 eV s^{-1} per H^+ ion were identified around the H^+ resonance velocity $V_{R_{H^+}}$ = ~870 to 1720 km s^{-1}, which was derived by using the dispersion relation (*16*). This energy loss is due to the agyrotropic distribution shown in Fig. 3, A to D. Because $|\mathbf{B}_{wave}|$ reached ~10% of $|\mathbf{B}_0|$ in this event, the observed wide extent of the interactions around $V_{R_{H^+}}$ is consistent (*16*) with the nonlinear trapping of H^+ by the large-amplitude cyclotron wave (*17*).

Shortly after the beginning of the wave (~12:18:24 UT), He^+ with a peak at ~1.5 keV was detected in ion composition data (Fig. 4, A and B). This coincides with an ion population observed with FPI-DIS in the corresponding energy range that is concentrated in pitch angle between 90° and 112.5° (Fig. 4, C and D). These ions were concentrated in less than four 11.25° gyro phase bins and were rotating with the wave—they were phase-bunched (Fig. 4, D and E). The maximum energy of He^+ (~3 keV) is nearly equal to those in the most energetic He^+ energization event (~2 keV) that have been reported in the magnetosphere (*19*) and ~10 times

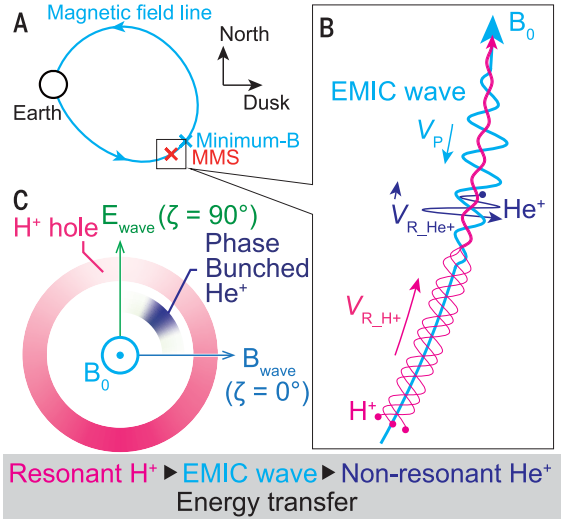


Fig. 1. Schematic diagrams summarizing EMIC wave propagation (direction of the phase velocity V_P), ion motion along the field line, and relative phase ζ distributions of He^+ and resonant H^+. (**A**) Positions of the MMS spacecraft and the point of lowest magnetic field (minimum-B). (**B** and **C**) Schematic of (B) the observed interactions and (C) ζ distributions of the phase-bunched He^+ with a small parallel velocity opposite to the He^+ resonance velocity $V_{R_{He^+}}$ and resonant H^+ with a parallel velocity equal to the H^+ resonance velocity $V_{R_{H^+}}$. Directions of the wave magnetic field $\mathbf{B}_{wave}$ and wave electric field $\mathbf{E}_{wave}$ relative to the background magnetic field $\mathbf{B}_0$, which is directed out of the page, are also shown in (C).

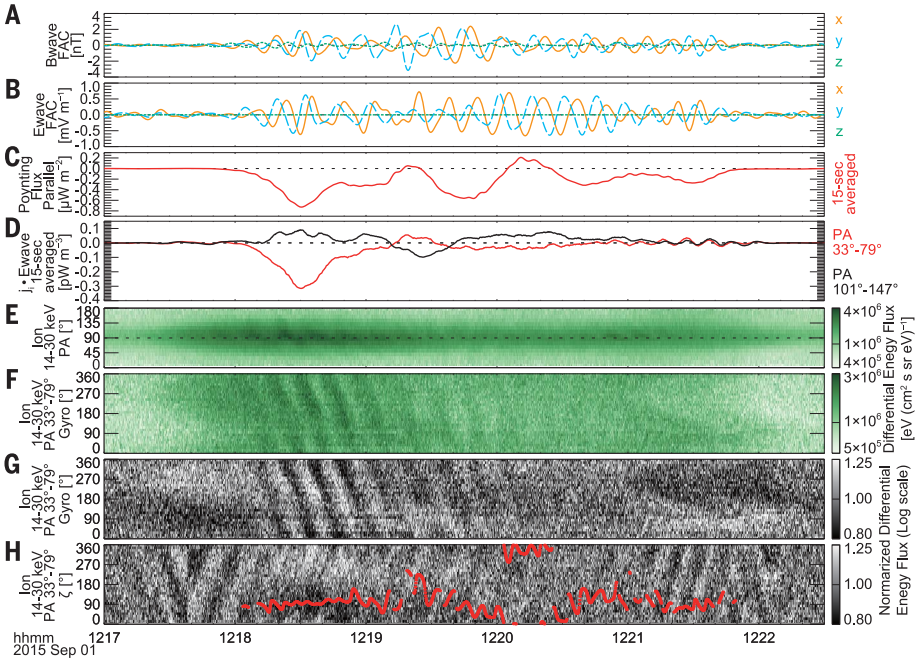


Fig. 2. Wave fields and hot ions (14 to 30 keV), averaged over all four spacecraft. (**A**) Wave magnetic field $\mathbf{B}_{wave}$. (**B**) Wave electric field $\mathbf{E}_{wave}$. (**C**) 15-s averaged Poynting flux **S** parallel to the background magnetic field. (**D**) 15-s averaged dot product of ion current $\mathbf{j}_i$ (14 to 30 keV) and $\mathbf{E}_{wave}$. (**E** and **F**) Pitch-angle (PA) and gyro phase spectrograms, respectively, of ion differential energy fluxes. (**G** and **H**) Gyro and relative phase (ζ) spectrograms, respectively, of normalized differential energy fluxes. Red dots mark the direction of $\mathbf{E}_{wave}$ when the amplitudes of $\mathbf{B}_{wave}$ and $\mathbf{E}_{wave}$ in the *x*-*y* plane of the field-aligned coordinate system were larger than 0.25 nT and 0.2 mV m^{-1}, respectively. The gyro and ζ spectrograms were constructed by using ions with pitch angles of 33.25° to 78.75°.

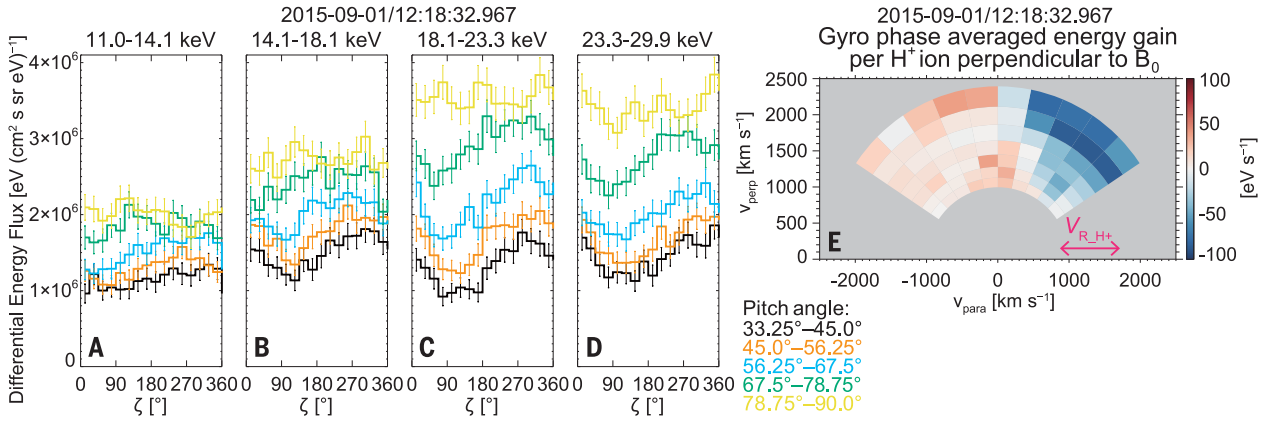


Fig. 3. Agyrotropy and energy gain of hot H^+. (**A** to **D**) Relative phase angle (ζ) distributions in 15-s averages of ion differential energy fluxes centered at 12:18:32.967 UT, split into pitch-angle bins from 33.25° to 90.0°. Distributions are separated by energy (A) 11.0 to 14.1 keV, (B) 14.1 to 18.1 keV, (C) 18.1 to 23.3 keV, and (D) 23.3 to 29.9 keV; error bars are 2σ. (**E**) Distribution for a 15-s gyro phase–averaged energy gain, perpendicular to the background magnetic field $\mathbf{B}_0$, per H^+ ion in the energy range of 5.1 to 29.9 keV in the velocity space: ion velocity parallel (v_{para}) and perpendicular (v_{perp}) to $\mathbf{B}_0$ plane.

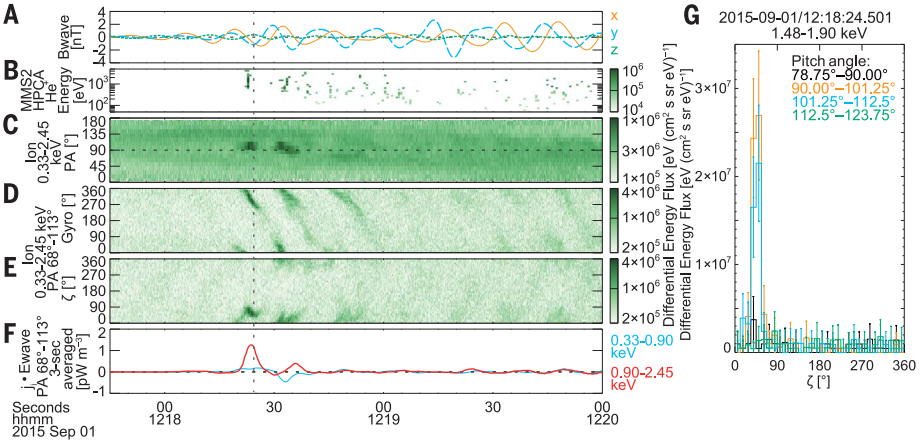


Fig. 4. Phase-bunched ions near the beginning of the wave event. (**A**) Wave magnetic field $\mathbf{B}_{wave}$. (**B**) Energy-time spectrogram of He^+. (**C** to **E**) Pitch-angle, gyro, and relative phase (ζ) spectrograms, respectively, of ion differential energy fluxes (0.33 to 2.45 keV). (**F**) $\mathbf{j}_i \cdot \mathbf{E}_{wave}$ averaged over 3 s. (**G**) ζ distributions of 1.05-s average of ion differential energy fluxes observed at 12:18:24.501 UT (vertical gray dashed line) by one spacecraft (MMS4) centered in pitch-angle bins from 78.75° to 123.75° and covering the energy bin 1.48 to 1.90 keV; error bars are 2σ. (C) and (D) were constructed by using ions with pitch angles of 67.5° to 112.5°.

higher than previous observation of bunched He^+ after a wave event (*20*). Positive $\mathbf{j}_i \cdot \mathbf{E}_{wave}$ around 12:18:24 UT (Fig. 4F) indicates that the He^+ ions with the highest flux in the event were being accelerated by $\mathbf{E}_{wave}$. In contrast to hot H^+ (Fig. 3, B to D), a sharp peak of ion fluxes appeared at ζ ~ 45°, which is ~45° from $\mathbf{E}_{wave}$ (ζ ~ 90°) (Figs. 1C and 4G and fig. S6). This provides evidence for an interaction in which almost all He^+ ions were accelerated by $\mathbf{E}_{wave}$, although energization of He^+ itself has been reported from the 1980s (*21*, *22*). The parallel motion of He^+ opposite to the direction of $V_{R_H e^+}$ is inconsistent with the cyclotron resonant acceleration, which has been considered as the most plausible candidate for He^+ energization perpendicular to $\mathbf{B}_0$ (*19*, *23*, *24*). Thus, the interaction must be of nonresonant type (*25*), a phenomenon that has not been simulated self-consistently. In cases in which the wave amplitude is large and the wave frequency is slightly different from the cyclotron frequency of the ion species, ions can be substantially accelerated over a time scale of approximately $\pi/|\Omega_i - \omega|$ because of slow rotation of the wave electric field as felt by the ions. Phase bunching is also predicted, if the ions are initially sufficiently cold. Simple test particle tracing by using the measured parameters can explain the observed pitch angle, accumulation in gyro phase, and most of the energization (fig. S7) (*16*).

Using MMS's high time-resolution measurements of ions with a full-sky field of view, together with composition-resolved ion measurements, we have quantitatively demonstrated the simultaneous occurrence of two concurrent energy transfers: one from hot anisotropic H^+ (the free-energy source) to the ion cyclotron wave via cyclotron resonance and the other from the wave to He^+ via nonresonant interaction (Fig. 1). This provides direct quantitative evidence for collisionless energy transfer between distinct particle populations via wave-particle interactions. Such measurements, including information on the gyro phase of energetic charged particles relative to wave fields, provide the capability to unambiguously identify which types of wave-particle interaction are occurring.

REFERENCES AND NOTES

1. R. B. Horne *et al.*, *Nature* **437**, 227–230 (2005).
2. G. D. Reeves *et al.*, *Science* **341**, 991–994 (2013).
3. R. M. Thorne *et al.*, *Nature* **504**, 411–414 (2013).
4. T. Chang *et al.*, *Geophys. Res. Lett.* **13**, 636–639 (1986).
5. V. Jordanova *et al.*, *J. Geophys. Res.* **111** (A11), A11S10 (2006).
6. D. Mourenas, A. V. Artemyev, Q. Ma, O. V. Agapitov, W. Li, *Geophys. Res. Lett.* **43**, 4155–4163 (2016).
7. Y. Y. Shprits, A. Kellerman, N. Aseev, A. Y. Drozdov, I. Michaelis, *Geophys. Res. Lett.* **44**, 1204–1209 (2017).
8. N. P. Meredith, R. B. Horne, R. R. Anderson, *J. Geophys. Res.* **106** (A7), 13,165–13,178 (2001).
9. Y. Miyoshi *et al.*, *J. Geophys. Res.* **108** (A1), 1004 (2003).
10. R. C. Allen *et al.*, *J. Geophys. Res. Space Phys.* **121**, 6458–6479 (2016).
11. M. P. Gough, P. J. Christiansen, K. Wilhelm, *J. Geophys. Res.* **95**, 12,287–12,294 (1990).
12. H. Fukuhara *et al.*, *Earth Planets Space* **61**, 765–778 (2009).
13. Y. Katoh *et al.*, *Ann. Geophys.* **31**, 503–512 (2013).
14. M. Shoji *et al.*, *Geophys. Res. Lett.* **44**, 8730–8738 (2017).
15. J. L. Burch, T. E. Moore, R. B. Torbert, B. L. Giles, *Space Sci. Rev.* **199**, 5–21 (2016).
16. Materials and methods are available as supplementary materials.
17. Y. Omura *et al.*, *J. Geophys. Res.* **115** (A7), A07234 (2010).
18. C. Pollock *et al.*, *Space Sci. Rev.* **199**, 331–406 (2016).
19. C. G. Mouikis *et al.*, *Geophys. Res. Lett.* **29**, 74-1 (2002).
20. B. H. Mauk, C. E. McIlwain, R. L. McPherron, *Geophys. Res. Lett.* **8**, 103–106 (1981).

21. D. T. Young *et al.*, *J. Geophys. Res.* **86** (A8), 6755–6772 (1981).
22. A. Roux *et al.*, *J. Geophys. Res.* **87** (A10), 8174–8190 (1982).
23. J. C. Zhang *et al.*, *J. Geophys. Res.* **115**, A06212 (2010).
24. J. C. Zhang *et al.*, *J. Geophys. Res.* **116** (A11), A11201 (2011).
25. J. Berchem, R. Gendrin, *J. Geophys. Res.* **90** (A11), 10,945–10,960 (1985).

ACKNOWLEDGMENTS

We thank the entire MMS team and instrument leads for data access and support. We acknowledge J.-A. Sauvaud, V. N. Coffey, J. C. Dorelli, L. A. Avanov, B. Lavraud, M. O. Chandler, and C. Schiff for their valuable roles in providing instrumentation and data production/quality for the Fast Plasma Investigation. We also gratefully acknowledge E. Grimes and the development team of the Space Physics Environment Data Analysis System (SPEDAS) software for their fruitful efforts in providing this software for our use. **Funding:** This research was supported by the NASA MMS Mission in association with NASA contract NNG04EB99C, and by Grants-in-Aid for Scientific Research (17H06140 to N.K. and Y.S.; 15H05747 to M.K., Y.M., and Y.K.; 15H05815 to M.S., Y.M., and Y.K.; 17K14402 to M.S.; 16H06286 to Y.M.; and 15H03730 to Y.K.) of the Japan Society for the Promotion of Science (JSPS). D.J.G. is supported by the NASA MMS Mission Guest Investigators program. IRAP contributions to Fast Plasma Investigation on MMS were supported by CNES and CNRS. Support for R.J.S.'s effort was provided under subcontract with the University of New Hampshire, in turn under contract from SwRI and NASA. **Author contributions:** N.K. conceived and designed the study, found this event from the data, analyzed the data, and wrote the initial draft. M.K. and M.S. contributed to the design of the study and interpretation of the results. Y.M. oversaw the production of the dataset and discussed its interpretation. H.H. contributed to interpretation of the result and to writing. S.N. contributed to the design of the study and interpretation of the result and prepared Fig. 1. Y.K. contributed to interpretation of the results. Y.S. led Japanese contribution to the development of FPI-DIS on the MMS spacecraft used to make the plasma measurements. S.Y. contributed to the development of FPI. D.J.G. assisted with the interpretation and analysis of the high-resolution plasma data and with the preparation of the paper. A.F.V. contributed to the FPI analysis codes and to the interpretation of the results. B.L.G. led the calibration and operation of FPI and to the development of the scientific data products and is responsible for the data. T.E.M. led the design of FPI and its Instrument Data Processing Unit and consulted in their development, calibration, and operation and in the writing. W.R.P. contributed to the calibration and operation of FPI and to the development of the scientific data products. C.J.P. led the development of FPI. C.T.R. supervised the building of the magnetometer on MMS, its calibration and data processing, and assisted in the writing. R.J.S. was responsible for the in-flight calibration of the fluxgate magnetometers. S.A.F. led the development of the Hot Plasma Composition Analyzer on MMS and is responsible for the data. J.L.B. led the MMS mission and assisted in the writing. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The MMS data can be accessed from the MMS Science Data Center at https://lasp.colorado.edu/mms/sdc/public. We used the Level-2 data from the FGM survey (located in fgm/srvy/l2/scpot); FPI-DIS burst (fpi/brst/l2/dis-dist); EDP fast survey (edp/fast/l2/scpot); and HPCA burst (hpca/brst/l2/ion, hpca/brst/l2/moments), all from the period 12:15:28 to 12:24:00 UT on 1 September 2015. The Space Physics Environment Data Analysis System (SPEDAS) software used to download and analyze the data are available from http://themis.ssl.berkeley.edu/socware/bleeding_edge/spdsw_r24826_2018-03-02.zip. The Kyoto University Plasma Dispersion Analysis Package (KUPDAP) that was used to calculate the dispersion relation of the cyclotron wave is available from http://space.rish.kyoto-u.ac.jp/software.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6406/1000/suppl/DC1
Materials and Methods
Figs. S1 to S7
References (*26*–*35*)

5 September 2017; accepted 4 July 2018
10.1126/science.aap8730

OPTICAL COMPUTING

All-optical machine learning using diffractive deep neural networks

Xing Lin[1,2,3]*, Yair Rivenson[1,2,3]*, Nezih T. Yardimci[1,3], Muhammed Veli[1,2,3], Yi Luo[1,2,3], Mona Jarrahi[1,3], Aydogan Ozcan[1,2,3,4]†

Deep learning has been transforming our ability to execute advanced inference tasks using computers. Here we introduce a physical mechanism to perform machine learning by demonstrating an all-optical diffractive deep neural network (D^2NN) architecture that can implement various functions following the deep learning–based design of passive diffractive layers that work collectively. We created 3D-printed D^2NNs that implement classification of images of handwritten digits and fashion products, as well as the function of an imaging lens at a terahertz spectrum. Our all-optical deep learning framework can perform, at the speed of light, various complex functions that computer-based neural networks can execute; will find applications in all-optical image analysis, feature detection, and object classification; and will also enable new camera designs and optical components that perform distinctive tasks using D^2NNs.

Deep learning is one of the fastest-growing machine learning methods (*1*). This approach uses multilayered artificial neural networks implemented in a computer to digitally learn data representation and abstraction and to perform advanced tasks in a manner comparable or even superior to the performance of human experts. Recent examples in which deep learning has made major advances in machine learning include medical image analysis (*2*), speech recognition (*3*), language translation (*4*), and image classification (*5*), among others (*1*, *6*). Beyond some of these mainstream applications, deep learning methods are also being used to solve inverse imaging problems (*7*–*13*).

Here we introduce an all-optical deep learning framework in which the neural network is physically formed by multiple layers of diffractive surfaces that work in collaboration to optically perform an arbitrary function that the network can statistically learn. Whereas the inference and prediction mechanism of the physical network is all optical, the learning part that leads to its design is done through a computer. We term this framework a diffractive deep neural network (D^2NN) and demonstrate its inference capabilities through both simulations and experiments. Our D^2NN can be physically created by using several transmissive and/or reflective layers (*14*), where each point on a given layer either transmits or reflects the incoming wave, representing an artificial neuron that is connected to other neurons of the following layers through optical diffraction (Fig. 1A). In accordance with the Huygens-Fresnel principle, our terminology is based on each point on a given layer acting as a secondary source of a wave, the amplitude and phase of which are determined by the product of the input wave and the complex-valued transmission or reflection coefficient at that point [see (*14*) for an analysis of the waves within a D^2NN]. Therefore, an artificial neuron in a D^2NN is connected to other neurons of the following layer through a secondary wave modulated in amplitude and phase by both the input interference pattern created by the earlier layers and the local transmission or reflection coefficient at that point. As an analogy to standard deep neural networks (Fig. 1D), one can consider the transmission or reflection coefficient of each point or neuron as a multiplicative "bias" term, which is a learnable network parameter that is iteratively adjusted during the training process of the diffractive network, using an error back-propagation method. After this numerical training phase, the D^2NN design is fixed and the transmission or reflection coefficients of the neurons of all layers are determined. This D^2NN design—once physically fabricated using techniques such as 3D-printing or lithography—can then perform, at the speed of light, the specific task for which it is trained, using only optical diffraction and passive optical components or layers that do not need power, thereby creating an efficient and fast way of implementing machine learning tasks.

In general, the phase and amplitude of each neuron can be learnable parameters, providing a complex-valued modulation at each layer, which improves the inference performance of the diffractive network (fig. S1) (*14*). For coherent transmissive networks with phase-only modulation, each layer can be approximated as a thin optical element (Fig. 1). Through deep learning, the phase values of the neurons of each layer of the diffractive network are iteratively adjusted (trained) to perform a specific function by feeding training data at the input layer and then computing the network's output through optical diffraction. On the basis of the calculated error with respect to the target output, determined by the desired function, the network structure and its neuron phase values are optimized via an error back-propagation algorithm, which is based on the stochastic gradient descent approach used in conventional deep learning (*14*).

To demonstrate the performance of the D^2NN framework, we first trained it as a digit classifier to perform automated classification of handwritten digits, from 0 to 9 (Figs. 1B and 2A). For this task, phase-only transmission masks were designed by training a five-layer D^2NN with 55,000 images (5000 validation images) from the MNIST (Modified National Institute of Standards and Technology) handwritten digit database (*15*). Input digits were encoded into the amplitude of the input field to the D^2NN, and the diffractive network was trained to map input digits into 10 detector regions, one for each digit. The classification criterion was to find the detector with the maximum optical signal, and this was also used as a loss function during the network training (*14*).

After training, the design of the D^2NN digit classifier was numerically tested using 10,000 images from the MNIST test dataset (which were not used as part of the training or validation image sets) and achieved a classification accuracy of 91.75% (Fig. 3C and fig. S1). In addition to the classification performance of the diffractive network, we also analyzed the energy distribution observed at the network output plane for the same 10,000 test digits (Fig. 3C), the results of which clearly demonstrate that the diffractive network learned to focus the input energy of each handwritten digit into the correct (i.e., the target) detector region, in accord with its training. With the use of complex-valued modulation and increasing numbers of layers, neurons, and connections in the diffractive network, our classification accuracy can be further improved (figs. S1 and S2). For example, fig. S2 demonstrates a Lego-like physical transfer learning behavior for D^2NN framework, where the inference performance of an already existing D^2NN can be further improved by adding new diffractive layers—or, in some cases, by peeling off (i.e., discarding) some of the existing layers—where the new layers to be added are trained for improved inference (coming from the entire diffractive network: old and new layers). By using a patch of two layers added to an existing and fixed D^2NN design (N = 5 layers), we improved our MNIST classification accuracy to 93.39% (fig. S2) (*14*); the state-of-the-art convolutional neural network performance has been reported as 99.60 to 99.77% (*16*–*18*). More discussion on reconfiguring D^2NN designs is provided in the supplementary materials (*14*).

Following these numerical results, we 3D-printed our five-layer D^2NN design (Fig. 2A), with each layer having an area of 8 cm by 8 cm, followed by 10 detector regions defined at the output plane of the diffractive network (Figs. 1B and 3A). We then used continuous-wave illumination at 0.4 THz to test the network's inference performance (Figs. 2, C and D). Phase values of

[1]Department of Electrical and Computer Engineering, University of California, Los Angeles, CA 90095, USA. [2]Department of Bioengineering, University of California, Los Angeles, CA 90095, USA. [3]California NanoSystems Institute (CNSI), University of California, Los Angeles, CA 90095, USA. [4]Department of Surgery, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: ozcan@ucla.edu

each layer's neurons were physically encoded using the relative thickness of each 3D-printed neuron. Numerical testing of this five-layer D^{2}NN design achieved a classification accuracy of 91.75% over ~10,000 test images (Fig. 3C). To quantify the match between these numerical testing results and our experiments, we 3D-printed 50 handwritten digits (five different inputs per digit), selected among the same 91.75% of the test images for which numerical testing was successful. For each input object that is uniformly illuminated with the terahertz source, we imaged the output plane of the D^{2}NN to map the intensity distribution for each detector region that is assigned to a digit. The results (Fig. 3B) demonstrate the success of the 3D-printed diffractive neural network and its inference capability: The average intensity distribution at the output plane of the network for each input digit clearly reveals that the 3D-printed D^{2}NN was able to focus the input energy of the beam and achieve a maximum signal at the corresponding detector region assigned for that digit. Despite 3D-printing errors, possible alignment issues, and other experimental error sources in our setup (*14*), the match between the experimental and numerical testing of our five-layer D^{2}NN design was found to be 88% (Fig. 3B). This relatively small reduction in the performance of the experimental network compared to our numerical testing is especially pronounced for the digit 0 because it is challenging to 3D-print the large void region at the center of the digit. Similar printing challenges were also observed for other digits that have void regions; e.g., 6, 8, and 9 (Fig. 3B).

Next, we tested the classification performance of D^{2}NN framework with a more complicated image dataset—i.e., the Fashion-MNIST dataset (*19*), which includes 10 classes, each representing a fashion product (t-shirts, trousers, pullovers, dresses, coats, sandals, shirts, sneakers, bags, and ankle boots; see fig. S3 for sample images). In general, for a coherently illuminated D^{2}NN, we can use the amplitude and/or phase channels of the input plane to represent data to be classified or processed. In our digit classification results reported earlier, input objects were encoded by using the amplitude channel, and to demonstrate the utility of the phase channel of the network input, we encoded each input image corresponding to a fashion product as a phase-only object modulation (*14*). Our D^{2}NN inference results (as a function of the number of layers, neurons, and connections) for classification of fashion products are summarized in figs. S4 and S5. To provide an example of our performance, a phase-only and a complex-valued modulation D^{2}NN with $N = 5$ diffractive layers (sharing the same physical network dimensions as the digit classification D^{2}NN shown in Fig. 2A) reached an accuracy of 81.13 and 86.33%, respectively (fig. S4). By increasing the number of diffractive layers to $N = 10$ and the total number of neurons to 0.4 million, our classification accuracy increased to 86.60% (fig. S5). For convolutional neural net–based standard deep learning, the state-of-the-art performance for Fashion-MNIST classification accuracy has been reported as 96.7%, using ~8.9 million learnable parameters and ~2.5 million neurons (*20*).

To experimentally demonstrate the performance of fashion product classification using a physical D^{2}NN, we 3D-printed our phase-only five-layer design and 50 fashion products used as test objects (five per class) on the basis of the same procedures employed for the digit classification diffractive network (Figs. 2A and 3), except that each input object information was encoded in the phase channel. Our results are summarized in Fig. 4, revealing a 90% match between the experimental and numerical testing of our five-layer D^{2}NN design, with five errors out of 50 fashion products. Compared with digit classification (six errors out of 50 digits; Fig. 3), this experiment yielded a slightly better match

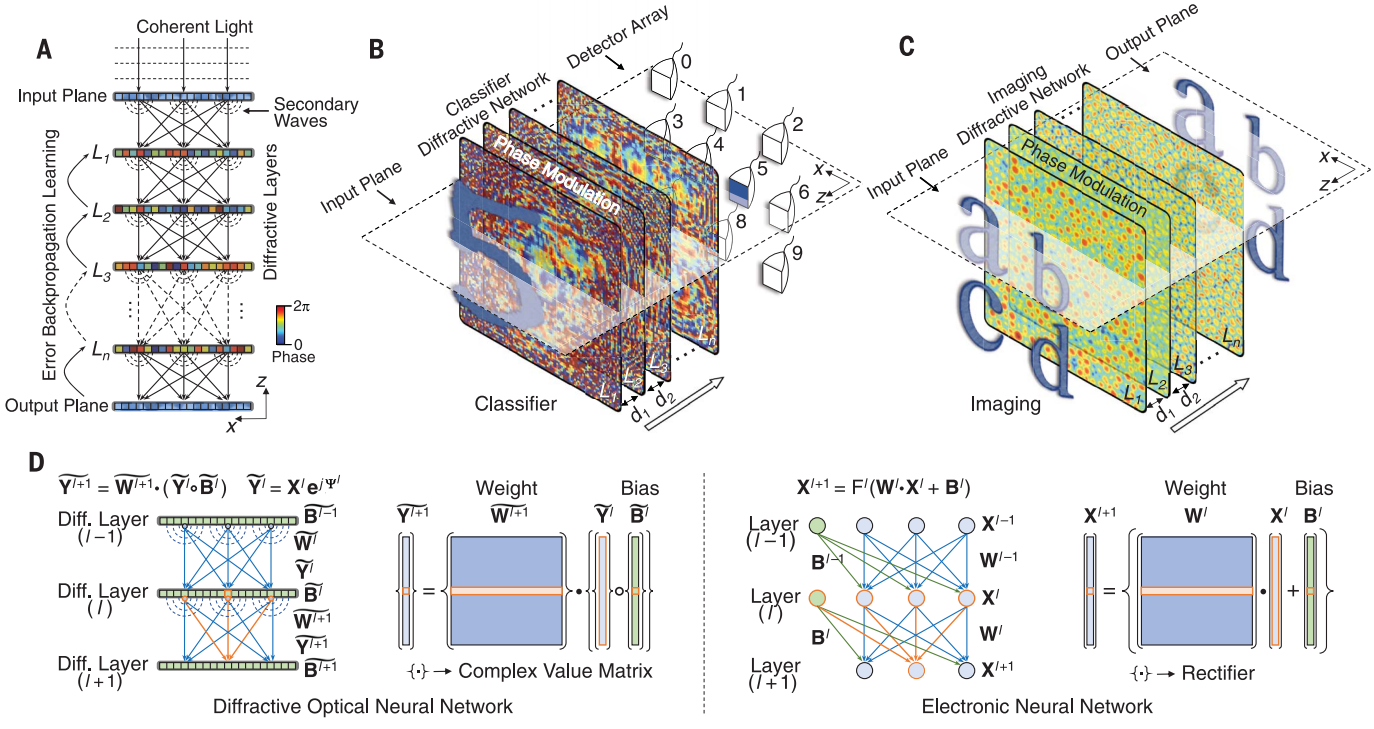


Fig. 1. Diffractive deep neural networks (D^{2}NNs). (**A**) A D^{2}NN comprises multiple transmissive (or reflective) layers, where each point on a given layer acts as a neuron, with a complex-valued transmission (or reflection) coefficient. The transmission or reflection coefficients of each layer can be trained by using deep learning to perform a function between the input and output planes of the network. After this learning phase, the D^{2}NN design is fixed; once fabricated or 3D-printed, it performs the learned function at the speed of light. *L*, layer. (**B** and **C**) We trained and experimentally implemented different types of D^{2}NNs: (B) classifier (for handwritten digits and fashion products) and (C) imager. *d*, distance. (**D**) Comparison between a D^{2}NN and a conventional neural network (*14*). Based on coherent waves, the D^{2}NN operates on complex-valued inputs, with multiplicative bias terms. Weights in a D^{2}NN are based on free-space diffraction and determine the interference of the secondary waves that are phase- and/or amplitude-modulated by the previous layers. "$\circ$" denotes a Hadamard product operation. "Electronic neural network" refers to the conventional neural network virtually implemented in a computer. **Y**, optical field at a given layer; Ψ, phase of the optical field; **X**, amplitude of the optical field; F, nonlinear rectifier function [see (*14*) for a discussion of optical nonlinearity in D^{2}NN].

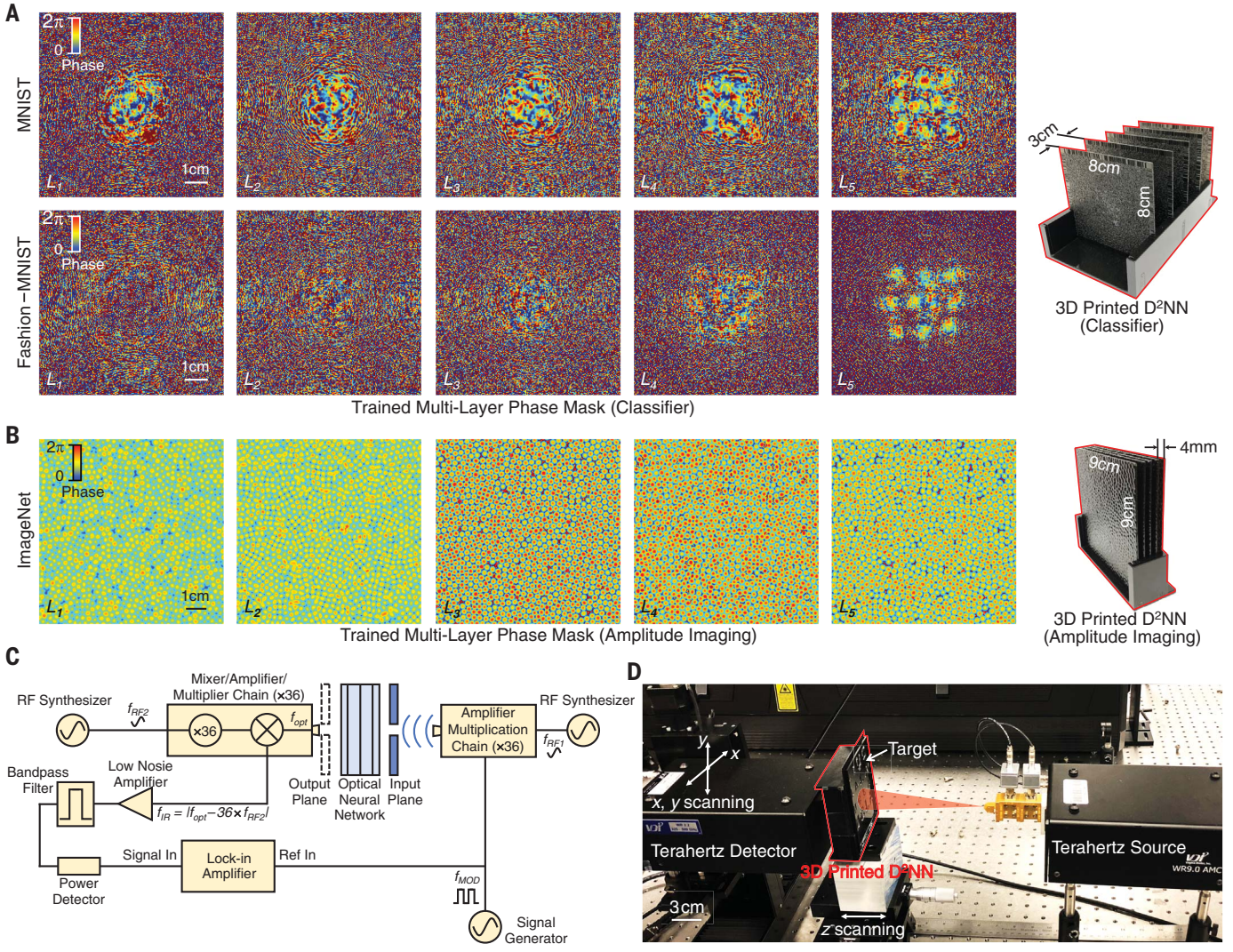


Fig. 2. Experimental testing of 3D-printed D²NNs. (**A** and **B**) After the training phase, the final designs of five different layers (L_1, L_2, ..., L_5) of the handwritten digit classifier, fashion product classifier, and the imager D²NNs are shown. To the right of the network layers, an illustration of the corresponding 3D-printed D²NN is shown. (**C** and **D**) Schematic (C) and photo (D) of the experimental terahertz setup. An amplifier-multiplier chain was used to generate continuous-wave radiation at 0.4 THz, and a mixer-amplifier-multiplier chain was used for the detection at the output plane of the network. RF, radio frequency; *f*, frequency.

between the experimental and numerical testing results (despite the more challenging nature of Fashion-MNIST dataset), perhaps because we used the phase channel, which does not suffer from the challenges associated with 3D-printing of void regions [such as in digits 0, 6, 8, and 9 (Fig. 3)], to encode input image information for fashion products.

Next, we tested the performance of a phase-only D²NN, composed of five 3D-printed transmission layers to implement amplitude imaging (Fig. 2B). This network was trained using the ImageNet database (*21*) to create a unit-magnification image of the input optical field amplitude at its output plane (~9 cm by 9 cm)—that is, the output image has the same physical size as the input object (*14*). As illustrated in fig. S6, A and C, the trained network initially connects every amplitude point at the input plane to various neurons and features of the following layers, which then focus the light back to a point at the output (i.e., image) plane, which is, as expected, quite different than the case of free-space diffraction (i.e., without the presence of the diffractive network), illustrated in fig. S6, B and D.

After training and blind testing, which served to numerically prove the imaging capability of the network (figs. S6 and S7), we then 3D-printed this designed D²NN. Using the same experimental setup shown in Fig. 2, C and D, we imaged the output plane of the 3D-printed D²NN for various input objects that were uniformly illuminated by continuous-wave radiation at 0.4 THz. Figure S8 summarizes our experimental results achieved with this 3D-printed D²NN, which successfully projected unit-magnification images of the input patterns at the output plane of the network, learning the function of an imager, or a physical auto-encoder. To evaluate the point spread function of this D²NN, we imaged pinholes with different diameters (1, 2, and 3 mm), which resulted in output images, each with a full width at half maximum of 1.5, 1.4, and 2.5 mm, respectively (fig. S8B). Our results also revealed that the printed network can resolve a linewidth of 1.8 mm at 0.4 THz (corresponding to a wavelength of 0.75 mm in air), which is slightly worse in resolution compared with the numerical testing of our D²NN design, where the network could resolve a linewidth of ~1.2 mm (fig. S7C). This experimental degradation in the performance of the diffractive network can be due to factors such as 3D-printing errors, potential misalignments, and absorption-related losses in the 3D-printed network (*14*).

Optical implementation of machine learning in artificial neural networks is promising because of the parallel computing capability and power efficiency of optical systems (*22*–*24*). Compared

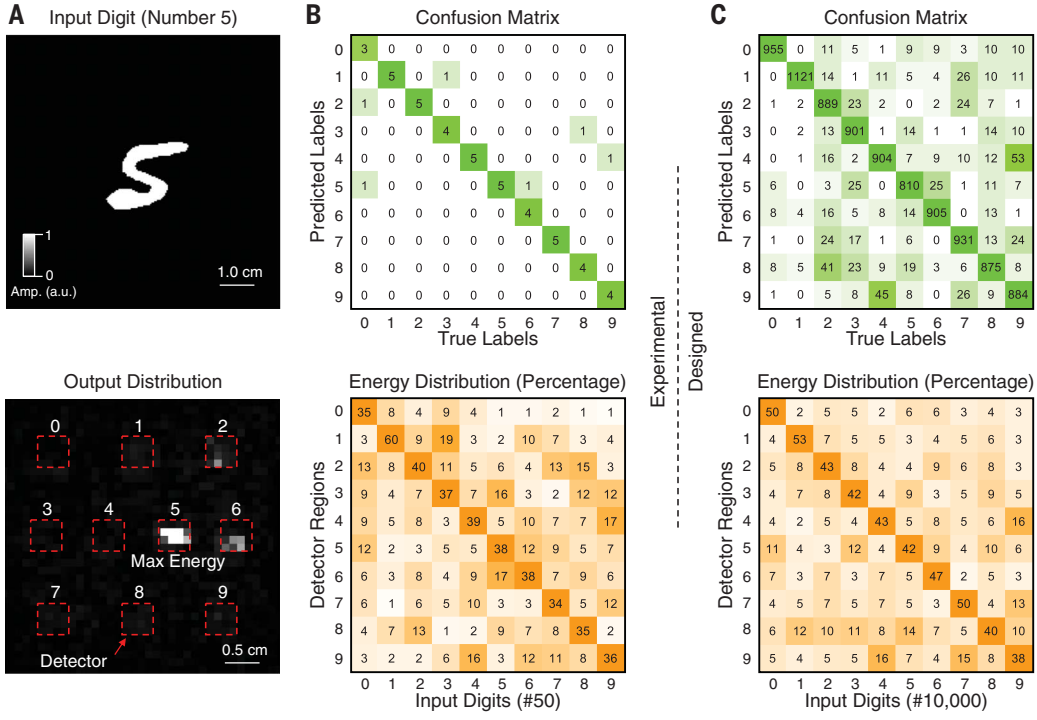


Fig. 3. Handwritten digit classifier D²NN. (**A**) A 3D-printed D²NN successfully classifies handwritten input digits (0, 1, ..., 9) on the basis of 10 different detector regions at the output plane of the network, each corresponding to one digit. As an example, the output image of the 3D-printed D²NN for a handwritten input of "5" is demonstrated, where the red dashed squares represent the trained detector regions for each digit. Other examples of our experimental results are shown in fig. S9. (**B**) Confusion matrix and energy distribution percentage for our experimental results, using 50 different handwritten digits (five for each digit) that were 3D-printed, selected among the images for which numerical testing was successful. (**C**) Same as (B), except summarizing our numerical testing results for 10,000 different handwritten digits (~1000 for each digit), achieving a classification accuracy of 91.75% using a five-layer design. Our classification accuracy increased to 93.39% by increasing the number of diffractive layers to seven, using a patch of two additional diffractive layers added to an existing and fixed D²NN (fig. S2).

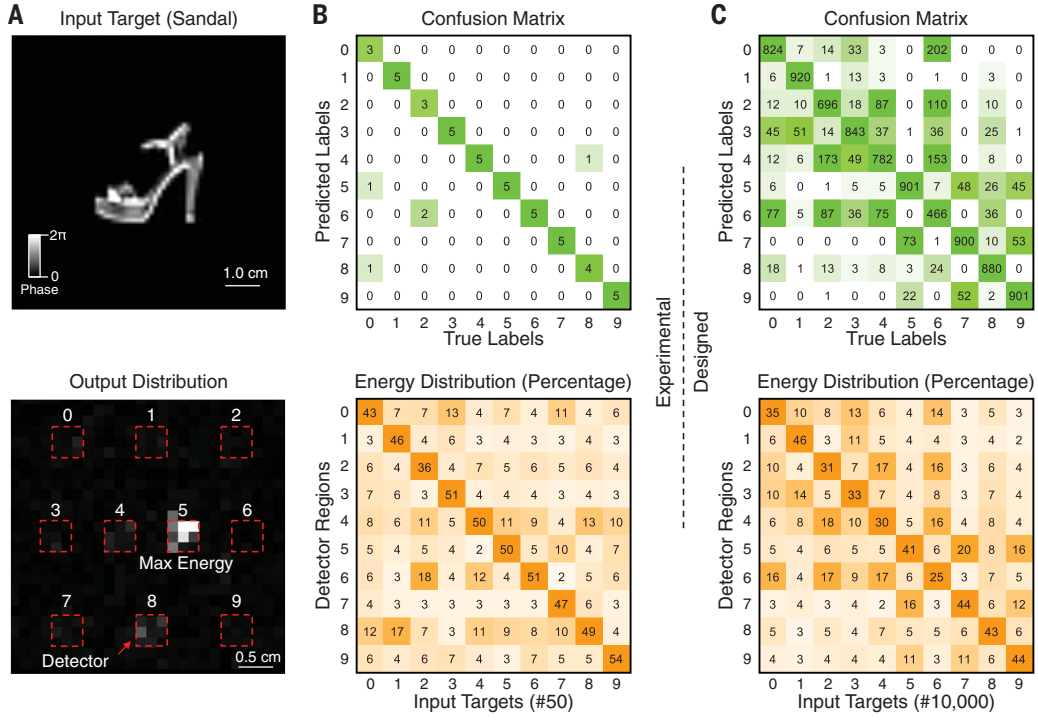


Fig. 4. Fashion product classifier D²NN. (**A**) As an example, the output image of the 3D-printed D²NN for a sandal input (Fashion-MNIST class 5) is demonstrated. The red dashed squares represent the trained detector regions for each fashion product. Other examples of our experimental results are shown in fig. S10. (**B**) Confusion matrix and energy distribution percentage for our experimental results, using 50 different fashion products (five per class) that were 3D-printed, selected among the images for which numerical testing was successful. (**C**) Same as (B), except summarizing our numerical testing results for 10,000 different fashion products (~1000 per class), achieving a classification accuracy of 81.13% using a five-layer design. By increasing the number of diffractive layers to 10, our classification accuracy increased to 86.60% (fig. S5).

with previous optoelectronics-based learning approaches (*22*, *25*–*27*), the D²NN framework provides a distinctive all-optical machine learning engine that efficiently operates at the speed of light using passive components and optical diffraction. An important advantage of D²NNs is that they can be easily scaled up using various high-throughput and large-area 3D-fabrication methods (such as soft lithography and additive manufacturing), as well as wide-field optical components and detection systems, to cost-effectively reach tens to hundreds of millions of neurons and hundreds of billions of connections in a scalable and power-efficient manner. For example, integration of D²NNs with lensfree on-chip imaging systems (*28*, *29*) could provide extreme parallelism within a cost-effective and portable platform. Such large-scale D²NNs may

be transformative for various applications, including image analysis, feature detection, and object classification, and may also enable new microscope or camera designs that can perform specific imaging tasks using D^2NNs. To achieve these new technologies, nonlinear optical materials (*14*) and a monolithic D^2NN design that combines all layers of the network as part of a 3D-fabrication method would be desirable. Among other techniques, laser lithography based on two-photon polymerization (*30*) can provide solutions for creating such D^2NNs.

REFERENCES AND NOTES

1. Y. LeCun, Y. Bengio, G. Hinton, *Nature* **521**, 436–444 (2015).
2. G. Litjens *et al*., *Med. Image Anal.* **42**, 60–88 (2017).
3. A. Graves, A. Mohamed, G. Hinton, in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, 2013), pp. 6645–6649.
4. K. Cho *et al*., arXiv:1406.1078 [cs.CL] (3 June 2014).
5. A. Krizhevsky, I. Sutskever, G. E. Hinton, in *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger, Eds. (Curran Associates, 2012), pp. 1097–1105.
6. D. Silver *et al*., *Nature* **529**, 484–489 (2016).
7. U. S. Kamilov *et al*., *Optica* **2**, 517 (2015).
8. Y. Rivenson *et al*., *Optica* **4**, 1437 (2017).
9. K. H. Jin, M. T. McCann, E. Froustey, M. Unser, *IEEE Trans. Image Process.* **26**, 4509–4522 (2017).
10. Y. Rivenson, Y. Zhang, H. Gunaydin, D. Teng, A. Ozcan, *Light Sci. Appl.* **7**, 17141 (2018).
11. A. Sinha, J. Lee, S. Li, G. Barbastathis, *Optica* **4**, 1117 (2017).
12. K. Hammernik *et al*., *Magn. Reson. Med.* **79**, 3055–3071 (2018).
13. Y. Rivenson *et al*., *ACS Photonics* **5**, 2354–2364 (2018).
14. Materials and methods are available as supplementary materials.
15. Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, *Proc. IEEE* **86**, 2278–2324 (1998).
16. D. Ciregan, U. Meier, J. Schmidhuber, in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2012), pp. 3642–3649.
17. C.-Y. Lee, P. W. Gallagher, Z. Tu, in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics* (AISTATS, 2016), pp. 464–472.
18. M. A. Ranzato, C. Poultney, S. Chopra, Y. LeCun, in *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, J. C. Platt, D. Koller, Y. Singer, S. T. Roweis, Eds. (Curran Associates, 2007), pp. 1137–1144.
19. H. Xiao, K. Rasul, R. Vollgraf, arXiv:1708.07747 [cs.LG] (25 August 2017).
20. S. Zeng, B. Zhang, J. Gou, arXiv:1802.07589 [cs.CV] (21 February 2018).
21. ImageNet database, www.image-net.org.
22. Y. Shen *et al*., *Nat. Photonics* **11**, 441–446 (2017).
23. D. Psaltis, D. Brady, X.-G. Gu, S. Lin, *Nature* **343**, 325–330 (1990).
24. K. H. Wagner, in *Proceedings of Frontiers in Optics 2017* (OSA, 2017), FW2C.1.
25. B. J. Shastri *et al*., arXiv:1801.00016 [cs.ET] (29 December 2017).
26. M. Hermans, M. Burm, T. Van Vaerenbergh, J. Dambre, P. Bienstman, *Nat. Commun.* **6**, 6729 (2015).
27. D. Brunner, M. C. Soriano, C. R. Mirasso, I. Fischer, *Nat. Commun.* **4**, 1364 (2013).
28. A. Greenbaum *et al*., *Nat. Methods* **9**, 889–895 (2012).
29. A. Ozcan, E. McLeod, *Annu. Rev. Biomed. Eng.* **18**, 77–102 (2016).
30. M. Emons *et al*., *Opt. Mater. Express* **2**, 942–947 (2012).

ACKNOWLEDGMENTS

We thank D. Mengu, Z. Wei, X. Wang, and Y. Xiao of UCLA for assistance with coding. **Funding:** The Ozcan Group at UCLA acknowledges the support of the National Science Foundation and the Howard Hughes Medical Institute. **Author contributions:** A.O., X.L., and Y.R. conceived of the research; X.L., N.T.Y., Y.L., Y.R., M.V., and M.J. contributed to the experiments; X.L., N.T.Y., M.V., and Y.R. processed the data; A.O., X.L., M.V., N.T.Y., Y.R., Y.L., and M.J. prepared the manuscript; and A.O. initiated and supervised the research. **Competing interests:** A.O., X.L., and Y.R. are inventors of a patent application on D^2NNs. **Data and materials availability:** All data and methods are present in the main text and supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6406/1004/suppl/DC1
Materials and Methods
Figs. S1 to S16
References (*31*–*46*)

6 April 2018; accepted 12 July 2018
Published online 26 July 2018
10.1126/science.aat8084

MEMBRANES

Zeolitic imidazolate framework membranes made by ligand-induced permselectivation

Xiaoli Ma*†, Prashant Kumar†, Nitish Mittal†, Alexandra Khlyustova, Prodromos Daoutidis, K. Andre Mkhoyan, Michael Tsapatsis*

Zeolitic imidazolate framework (ZIF) membranes are emerging as a promising energy-efficient separation technology. However, their reliable and scalable manufacturing remains a challenge. We demonstrate the fabrication of ZIF nanocomposite membranes by means of an all-vapor-phase processing method based on atomic layer deposition (ALD) of ZnO in a porous support followed by ligand-vapor treatment. After ALD, the obtained nanocomposite exhibits low flux and is not selective, whereas after ligand-vapor (2-methylimidazole) treatment, it is partially transformed to ZIF and shows stable performance with high mixture separation factor for propylene over propane (an energy-intensive high-volume separation) and high propylene flux. Membrane synthesis through ligand-induced permselectivation of a nonselective and impermeable deposit is shown to be simple and highly reproducible and holds promise for scalability.

Substantial energy and capital cost savings are possible by using membrane-based separations (*1*). Among the membrane materials that have been explored, zeolitic imidazolate frameworks (ZIFs) (*2*, *3*), particularly ZIF-8 and ZIF-67 (*4*–*6*), exhibit promising performance for the processing of important industrial mixtures, such as propylene/propane (*7*), that are considered challenging to separate efficiently with the currently used distillation-based methods (*8*, *9*). ZIF-8 membranes have been fabricated as continuous thin films on the outer surface or inside porous ceramic (*10*–*12*) or polymeric supports (*13*, *14*) with solution-based methods that are difficult to scale up. Replacing all solvothermal steps with completely solvent-free processing has environmental, cost, and scale-up advantages and has been adapted successfully for seeded growth of zeolite membranes (*15*, *16*). Recently, an all-vapor deposition technique was reported for the fabrication of metal-organic framework (MOF) thin films on silicon wafers for microelectronic applications (*17*). It is based on a combination of oxide deposition by means of atomic layer deposition (ALD), followed by 2-methylimidazole (mIm) ligand-vapor treatment to convert the ALD-deposited oxide to ZIF. However, only aspects of this approach have been adapted for membrane preparation, and an all-vapor membrane synthesis method is yet to be demonstrated (*18*, *19*). Here, we report the liquid/gel-free and seed-free synthesis of high-performance membranes through ligand-induced permselectivation (LIPS) of a nonselective and impermeable deposit (Fig. 1A). Distinct from all prior methods that aim to create a thin selective MOF molecular sieve layer by gradually filling the pores of a support or by forming a deposit on its surface, the current method first blocks the pores with an impermeable deposit, which is then transformed by means of LIPS to a selective MOF. As a result, the inherent drawback associated with all pore-filling membrane growth methods, which leave behind nonselective transport pathways (such as pinholes and nonselective grain boundaries), is circumvented.

α-Alumina macroporous substrates coated with a ~5-μm γ-alumina mesoporous layer with pores in the 2 to 5 nm range (*20*) were used as the supports. They can be made easily based on well-known procedures and are amenable to scale up in disc or tubular geometries. They have propylene permeance of 10^{-6} mol Pa^{-1} m^{-2} s^{-1} and do not exhibit propylene/propane selectivity. During ALD, diethylzinc reacts with the surface hydroxyl groups that are present in the mesopores of the γ-alumina layer. Subsequent introduction of water vapor yields hydroxylated zinc oxide, completing an ALD cycle. This process is repeated for up to 50 cycles so as to obtain a zinc oxide and/or zinc hydroxide (called ZnO from now on) deposit on top and inside the γ-alumina layer. As shown by the open symbols in Fig. 1B, an initially gradual propylene permeance reduction is followed by a large drop after 10 ALD cycles of ZnO. Specifically, for up to eight ALD cycles, propylene permeance remained above 10^{-8} mol Pa^{-1} m^{-2} s^{-1}, whereas after 10 ALD cycles, the resulting ZnO-alumina composite is rendered essentially impermeable, with a propylene permeance falling by more than four orders of magnitude to 2×10^{-11} mol Pa^{-1} m^{-2} s^{-1}, indicating that the pores of the substrate are essentially blocked for propylene by the ALD deposit. The abrupt permeance drop is typical of percolation-based densification observed in dense membranes formed through chemical vapor deposition of oxides inside porous supports (*21*, *22*). At no point during ALD do the membranes exhibit propylene/propane selectivity (Fig. 1C, open triangles).

After exposing the impermeable and nonselective 10-cycle ALD–modified supports to vapors generated by the sublimation of mIm, propylene permeance and propylene/propane selectivity increase by three and two orders of magnitude, respectively, as indicated by the tie-lines with the upward pointing arrows in Fig. 1, B and C. ZnO deposits of more than 20 cycles result in membranes with lower permeance (Fig. 1B). Membranes obtained from ZnO made by less than 10 ALD cycles showed relatively low selectivity (5 to 20) and higher permeances of ~10^{-7} mol Pa^{-1} m^{-2} s^{-1} (Fig. 1, B and C). Although these higher-flux lower-selectivity membranes could be of interest, here we focus on 10 and 20 cycles of ALD that reproducibly yield membranes with high selectivity (~100) and good permeance (>10^{-8} mol Pa^{-1} m^{-2} s^{-1}). Because of the ligand-induced transformation from an impermeable to selective membrane, the process described here is named LIPS.

The overall performances of LIPS membranes, in terms of propylene permeance and propylene/propane selectivity, are among the best that have been reported for both ZIF-8 and ZIF-67 membranes (Fig. 1D and table S1) (*5*, *6*, *10*, *13*, *18*, *19*, *23*–*28*). Under an equimolar propylene/propane mixture feed, and pressure as high as ~7 atm, the membranes made by means of LIPS exhibit a combination of high selectivity/separation factor and high propylene flux (Fig. 1E). In addition, stable performance in mixed gas separation tests under a feed pressure from 1 to 7 atm was demonstrated (fig. S1). Moreover, heating the membrane under equimolar 1 atm propylene/propane feed to 60°C for ~36 hours did not alter the room-temperature membrane performance. At 7 atm feed, depending on feed composition and permeate conditions (vacuum and 1 atm undiluted permeate), typical propylene fluxes range from 0.01 to 0.06 mol m^{-2} s^{-1} with separation factors of ~50 to 70 (figs. S2 and S3). A conservative process-scale assessment (*29*) shows promise for large-scale uses such as de-bottlenecking of distillation columns in order to increase processing capacity.

To illuminate the membrane microstructure, reveal the location of the ZIF-8–selective layer, and relate it with that of the as-made and ALD-modified alumina support, we used x-ray diffraction (XRD) and electron microscopy imaging, coupled with gradual ZIF-8 removal through water washing. A high-angle annular dark-field scanning transmission electron microscopy (HAADF-STEM) image of a cross section of the porous support (γ-alumina supported on α-alumina) prepared by focused ion beam is shown in Fig. 2A along with Al and Zn spatial maps obtained with energy-dispersive x-ray (EDX) spectral imaging and the corresponding Al- and Zn-composition line scans. As expected, the support does not contain any Zn. The 5-μm uniform γ-alumina mesoporous layer is clearly distinguished from

Department of Chemical Engineering and Materials Science, University of Minnesota, 421 Washington Avenue SE, Minneapolis, MN 55455, USA.
***Corresponding author. Email: tsapatsis@umn.edu (M.T.); maxx0183@umn.edu (X.M.)** †These authors contributed equally to this work.

the underlying α-alumina macroporous support. The corresponding images and scans after 10 cycles of ALD (before mIm-vapor treatment) are shown in Fig. 2B. Zn is detected throughout the γ-alumina layer but is mostly present at its top 200 nm. There is no detectable Zn in the α-alumina. The line scans in Fig. 2B indicate that Zn is also present as a thin layer on the outside surface of the γ-alumina layer. Top-view scanning electron microscopy (SEM) imaging shows the presence of partially intergrown nanoparticles, which have distinct morphology compared with that of the fibrous appearance of the γ-alumina top surface (fig. S4, A and B) and can be attributed to ZnO that was deposited outside the pores of the support. The drastic reduction in propylene permeance, which essentially renders the membrane propylene-impermeable after 10 ALD cycles (Fig. 1), can be attributed to γ-alumina mesopore blockage by this thin (less than 0.5 μm) deposit that is partly confined in the mesopores and partly present on their external top surface. The corresponding set after mIm-vapor-treatment is shown in Fig. 2C; the Zn distribution changes substantially. More Zn is now present within the rest of the γ-alumina layer, and a new Zn-containing deposit at the γ-alumina/α-alumina interface is evident. The observations described above are more clearly visible in Fig. 2, D to F, which shows the corresponding higher-magnification cross-sectional images from the top and bottom parts of the γ-alumina before (Fig. 2D) and after ALD (Fig. 2E) and after mIm-vapor treatment (Fig. 2F).

The altered Zn profiles demonstrate that during mIm-vapor treatment, the ALD-deposited Zn can be mobilized and transported by diffusion throughout the 5-μm γ-alumina mesoporous layer. Consistently, nitrogen-composition line-scans indicate the presence of mIm throughout the γ-alumina layer, with higher concentrations at the top and bottom (fig. S5), which is coincident with the Zn content maxima. The pronounced redistribution of the deposited Zn suggests formation of Zn-mIm species with increased surface mobility compared with that of the ALD-deposited ZnO. In the geometry used for mIm-vapor treatment, mIm vapors are introduced from both sides of the γ-alumina layer (not only from its top side but also from the bottom γ-alumina/α-alumina interface), reacting with and mobilizing the deposited ZnO in the interior of the γ-alumina layer. The earlier report on mIm-vapor treatment of the ALD-deposited ZnO is also suggestive of Zn-mIm mobile species because the morphology of the ZIF-8 crystals formed upon exposure to mIm is coarsened with respect to that of the ALD-deposited oxide (*17*).

Considering the low ZIF density, if we assume full conversion to ZIF-8, the levels of Zn detected within the γ-alumina layer after mIm-vapor treatment are sufficient to fill the 2- to 5-nm pores and create a selective ZIF deposit. However, the high levels of Zn remaining near, and at the top, of the γ-alumina layer indicate that not all of the ALD-deposited ZnO has been transformed to ZIF after mIm-vapor treatment (there should be a substantial remaining unconverted ZnO fraction). This is further corroborated with

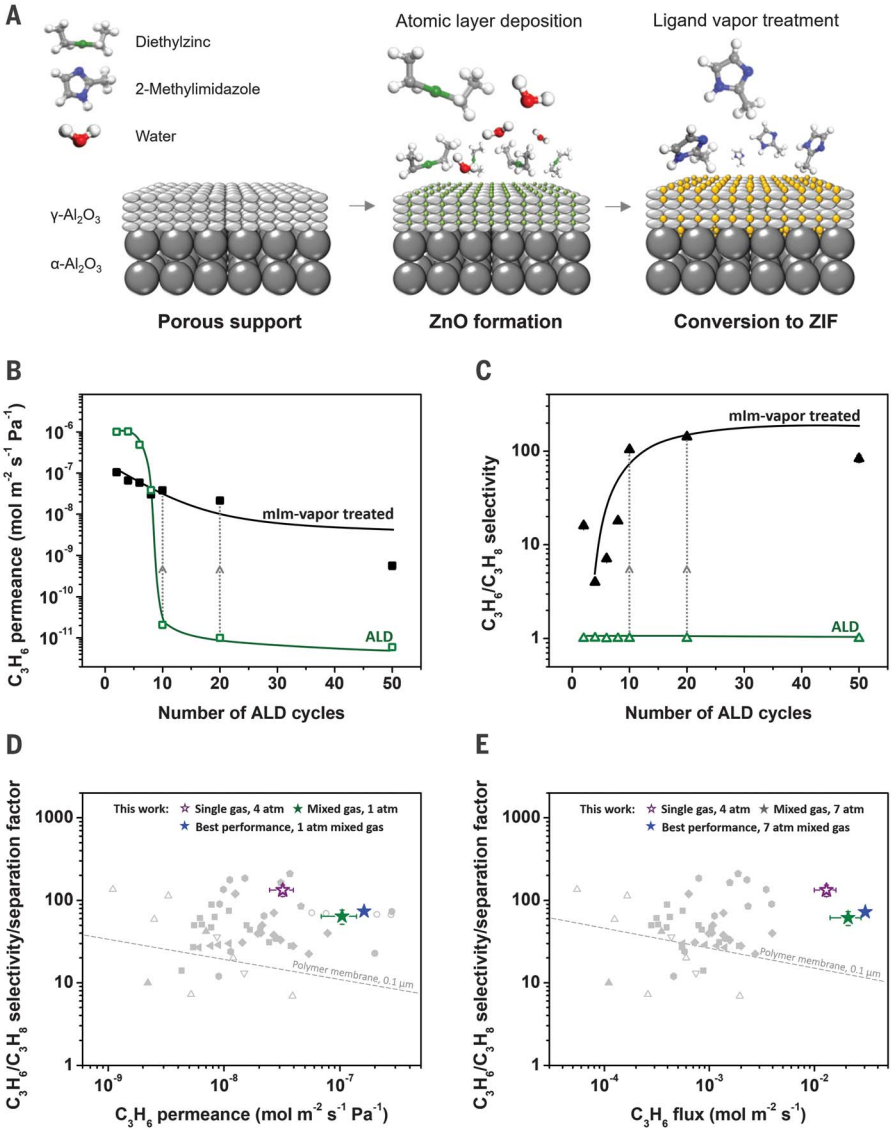


Fig. 1. Permeation properties of membranes made by LIPS. (**A**) Schematic of the all-vapor-phase LIPS membrane fabrication process. The pores of support are first blocked with ZnO made by means of ALD. The impermeable ZnO deposits are converted to ZIF by means of ligand-vapor treatment. (**B** and **C**) Propylene permeances (B) and propylene/propane single-component selectivities (C) of the ZIF-8/γ-alumina nanocomposite membranes as a function of the number of ALD cycles (solid symbols). Open symbols are values obtained from a support treated by the indicated cycles of ZnO ALD (before the ligand-vapor treatment). The tie lines connect them with the corresponding values obtained after the ligand-vapor treatment. (**D** and **E**) Propylene/propane single-component selectivities (open symbols) and equimolar mixture separation factors (solid symbols) versus propylene permeance (D) and versus propylene flux (E) obtained from the ZIF-8/γ-alumina nanocomposite membranes and as reported in the literature [solid square, (*5*, *23*); left-pointing solid triangle, (*24*);up-pointing open triangle, (*25*); up-pointing solid triangle, (*26*); down-pointing open triangle, (*27*); solid pentagon, (*6*); solid diamond, (*10*, *18*); solid circle, (*19*); solid hexagon, (*13*, *28*); and additional references cited in table S1]. The feed pressure for single-component permeation test is ~4 atm, and the feed pressure for mixed gas (equimolar feed) separation test is 1 atm in (D) and ~7 atm in (E). The performance data are the average values of six different nanocomposite membranes made from 10 and 20 cycles ZnO ALD. The lines represent Robeson upper bounds for polymeric membranes (*32*) assuming polymer thickness of 0.1 μm and, for the estimation of flux in (E), a 7-atm equimolar feed.

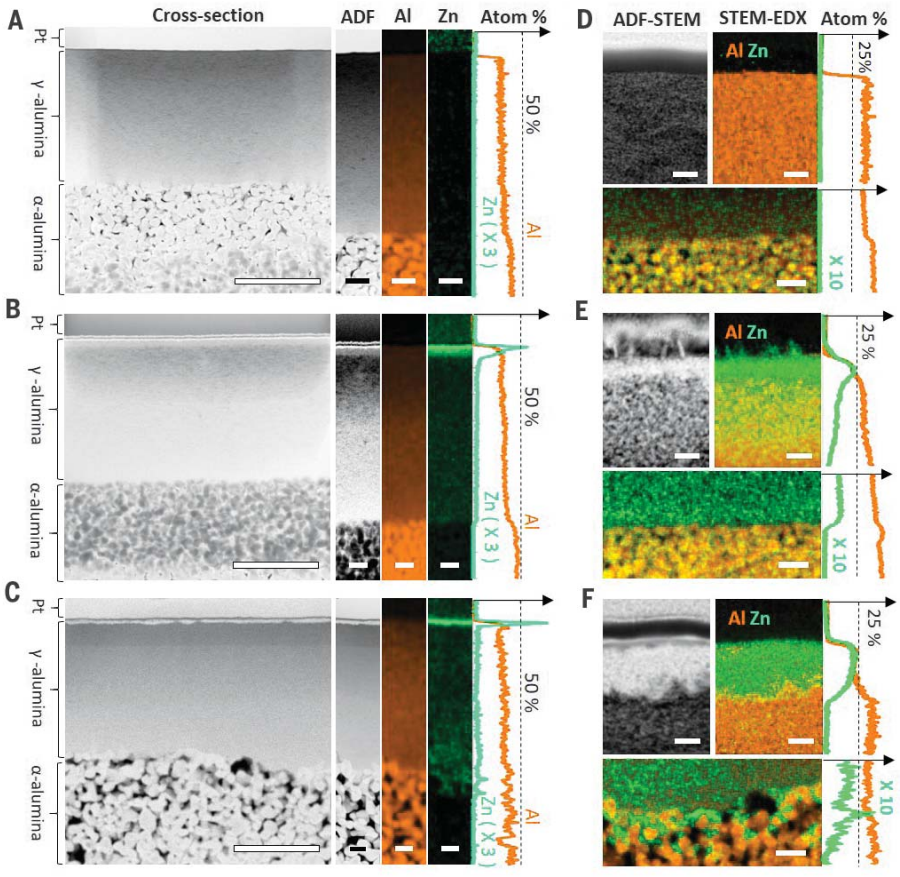


Fig. 2. Microstructure of membranes made by means of LIPS. (**A** to **C**) Cross section analysis with ADF-STEM imaging, corresponding spatial distribution of aluminum (orange) and zinc (green), and the averaged atomic percentage across the same section along the depth for (A) γ-alumina on α-alumina support, (B) after ZnO deposition by ALD, and (C) after mIm ligand-vapor treatment. Zinc atomic percentage has been magnified three times for clarity. (**D** to **F**) Magnified views of top (γ-Al_2O_3) and bottom (γ-Al_2O_3/α-Al_2O_3 interface) sections of (A) to (C), respectively. Zinc signal has been amplified 10 times for the γ-Al_2O_3/α-Al_2O_3 section for visibility. (A) to (C), large scale bars, 2 μm; small scale bars, 400 nm. (D) to (F), small scale bars (top), 50 nm; larger scale bars (bottom), 500 nm.

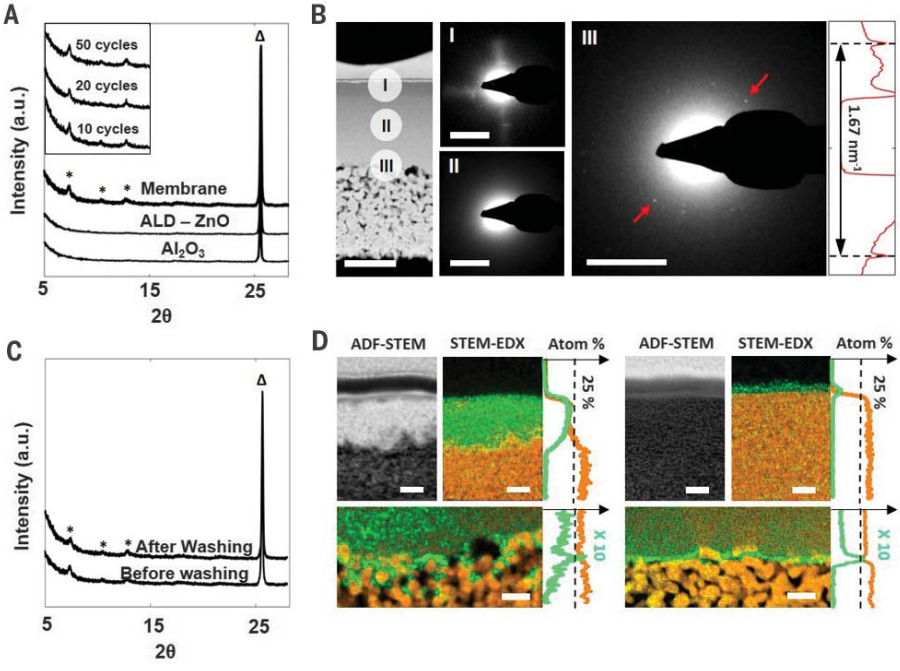


Fig. 3. Diffraction analysis of LIPS membranes before and after partial deposit removal by water washing. (**A**) XRD patterns of alumina support (5-μm-thick γ-alumina supported on α-alumina), of the nanocomposite obtained after 10 cycles of ZnO ALD, and of the ZIF-8/γ-alumina nanocomposite membrane obtained after ligand-vapor treatment of the ZnO deposit made with 10 ALD cycles. (Inset) Magnified views of the XRD patterns (2θ range of ZIF-8) of membranes prepared after ligand-vapor treatment of the ZnO deposit made by 10, 20, and 50 ALD cycles. Asterisks and open triangles indicate ZIF-8 and α-alumina peaks, respectively. (**B**) Selected area ED patterns acquired at different depths as marked in the cross section. Diffraction patterns I and II show no evidence of crystallinity, whereas pattern III shows sharp spots. Line scans across the spots indicated with red arrows show interplanar distances of 1.2 nm (0.83 nm^{-1}) corresponding to (011) spacing for ZIF-8 structure. Scale bars, 1 nm^{-1}. (**C**) Diffraction patterns acquired from a ZIF membrane before and after washing. Asterisks and open triangles indicate ZIF-8 and α-alumina peaks, respectively. (**D**) ADF-STEM image and the corresponding spatial elemental distribution of aluminum (orange) and zinc (green) in (top) γ-Al_2O_3 and (bottom) γ-Al_2O_3/α-Al_2O_3 interface of (left) ZIF membrane before washing and (right) ZIF membrane after washing. The zinc signal has been amplified 10 times for the γ-Al_2O_3/α-Al_2O_3 section for visibility. Scale bars, top, 50 nm; bottom, 500 nm.

top-view SEM imaging (fig. S4C) that shows preservation of the nanoparticulate microstructure with enlarged grain appearance that can be attributed to, at most, a partial transformation of the ALD deposit to ZIF.

We attempted to detect the presence of ZIF-8 with XRD and electron diffraction (ED). After imidazole-vapor treatment, XRD peaks that are characteristic of ZIF-8 can be clearly detected (Fig. 3A). Examination of the membrane cross section by means of ED (Fig. 3B) shows that ZIF-8 is detectable only at the γ-alumina/α-alumina interface. In the other regions of the γ-alumina layer, crystalline ZIF-8 could not be detected despite the presence of Zn. This finding raised the question of whether the deposit at the γ-alumina/α-alumina interface is solely responsible for the selective membrane performance.

To answer this question, the γ-alumina top side of a membrane was washed with flowing deionized (DI) water in order to gradually dissolve the ZIF-8 or ZIF-8–like deposits and alter its permeation performance and microstructure. In a control experiment, it was determined that washing with water, a 10-ALD-cycle membrane (before imidazole-vapor treatment) did not alter its characteristic, propylene-impermeable performance. By contrast, the permeation performance of the ZIF-8 membrane changed upon washing. After a first washing with 100 g of DI water, the propylene permeance of the membrane tested increased from 170 to 195 gas permeation units (GPU), while the mixture separation factor decreased from 45 to 5.6. After a second washing, the propylene permeance increased to 329 GPU, and the separation factor dropped to 1.7. XRD of the membrane after the second washing did not reveal substantial changes in the ZIF-8 reflections (Fig. 3C). By contrast, cross-sectional examination by means of HAADF-STEM and spectral imaging (Fig. 3D) revealed that the Zn present within the γ-alumina layer decreased, whereas the ZIF-8 present at the γ-alumina/α-alumina interface was largely preserved.

These results reveal that the deposit in the γ-alumina/α-alumina interface is the major contributor to the detectable XRD ZIF-8 reflections, but the propylene-selective performance of the membrane is mostly due to the deposit within and on top of the γ-alumina layer. The lack of detectable crystallinity in the selective layer may be attributed to its confinement within the small 2- to 5-nm pores of the γ-alumina layer and in between the ZnO grains of the top deposit. The high separation factors achieved should then be attributed mostly to this not well-crystallized ZIF deposit and could be explained by the recently identified structural similarities of ZIF-8 and amorphous ZIFs (*30*, *31*).

The LIPS process, demonstrated here, establishes a reliable, scalable, and robust approach for the fabrication of ZIF and possibly other MOF membranes and nanocomposites. Unlike other molecular sieve membrane fabrication methods, which rely on solvothermal nucleation and growth that is difficult to reliably scale up, the method shown here is based on scalable, solvent-free, seed-free, all-vapor processing with ALD, a well-established materials processing technology.

REFERENCES AND NOTES

1. D. S. Sholl, R. P. Lively, *Nature* **532**, 435–437 (2016).
2. K. S. Park *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 10186–10191 (2006).
3. H. Furukawa, K. E. Cordova, M. O'Keeffe, O. M. Yaghi, *Science* **341**, 1230444 (2013).
4. H. Bux *et al.*, *J. Am. Chem. Soc.* **131**, 16000–16001 (2009).
5. Y. C. Pan, T. Li, G. Lestari, Z. P. Lai, *J. Membr. Sci.* **390-391**, 93–98 (2012).
6. H. T. Kwon, H. K. Jeong, A. S. Lee, H. S. An, J. S. Lee, *J. Am. Chem. Soc.* **137**, 12304–12311 (2015).
7. K. Li *et al.*, *J. Am. Chem. Soc.* **131**, 10368–10369 (2009).
8. R. B. Eldridge, *Ind. Eng. Chem. Res.* **32**, 2208–2212 (1993).
9. M. Galizia *et al.*, *Macromolecules* **50**, 7809–7843 (2017).
10. M. J. Lee, H. T. Kwon, H. K. Jeong, *Angew. Chem. Int. Ed.* **57**, 156–161 (2018).
11. E. Jang *et al.*, *J. Membr. Sci.* **540**, 430–439 (2017).
12. M. Drobek *et al.*, *J. Membr. Sci.* **475**, 39–46 (2015).
13. A. J. Brown *et al.*, *Science* **345**, 72–75 (2014).
14. S. C. Hess, R. N. Grass, W. J. Stark, *Chem. Mater.* **28**, 7638–7644 (2016).
15. T. C. T. Pham, T. H. Nguyen, K. B. Yoon, *Angew. Chem. Int. Ed.* **52**, 8693–8698 (2013).
16. M. Y. Jeon *et al.*, *Nature* **543**, 690–694 (2017).
17. I. Stassen *et al.*, *Nat. Mater.* **15**, 304–310 (2016).
18. H. T. Kwon *et al.*, *Chem. Commun.* **52**, 11669–11672 (2016).
19. W. Li *et al.*, *Nat. Commun.* **8**, 406 (2017).
20. C. H. Chang, R. Gopalan, Y. S. Lin, *J. Membr. Sci.* **91**, 27–45 (1994).
21. Y. S. Lin, A. J. Burggraaf, *Chem. Eng. Sci.* **46**, 3067–3080 (1991).
22. M. Tsapatsis, G. R. Gavalas, *AIChE J.* **38**, 847–856 (1992).
23. Y. C. Pan, W. Liu, Y. J. Zhao, C. Q. Wang, Z. P. Lai, *J. Membr. Sci.* **493**, 88–96 (2015).
24. D. F. Liu, X. L. Ma, H. X. Xi, Y. S. Lin, *J. Membr. Sci.* **451**, 85–93 (2014).
25. N. Hara *et al.*, *J. Membr. Sci.* **450**, 215–223 (2014).
26. N. Hara *et al.*, *J. Jpn. Petrol. Inst.* **58**, 237–244 (2015).
27. S. Tanaka, K. Okubo, K. Kida, M. Sugita, T. Takewaki, *J. Membr. Sci.* **544**, 306–311 (2017).
28. K. Eum *et al.*, *Adv. Funct. Mater.* **26**, 5011–5018 (2016).
29. Materials and methods are available as supplementary materials.
30. P. Adhikari *et al.*, *J. Phys. Chem. C* **120**, 15362–15368 (2016).
31. H. Tao, T. D. Bennett, Y. Yue, *Adv. Mater.* **29**, 1601705 (2017).
32. W. J. Koros, R. P. Lively, *AIChE J.* **58**, 2624–2633 (2012).

ACKNOWLEDGMENTS

The authors acknowledge R. Agrawal and his group (Purdue University) for a valuable discussion. **Funding:** This work was supported by the Center for Gas Separations Relevant to Clean Energy Technologies, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences under award DE-SC0001015. Parts of this work were carried out in the Characterization Facility, University of Minnesota and the Minnesota Nano Center, which receive partial support from the National Science Foundation through the Materials Research Science and Engineering Center and National Nanotechnology Coordinated Infrastructure programs, respectively. SEM measurements were partially performed on a Hitachi 8230 provided by NSF MRI DMR-1229263. **Author contributions:** M.T. and X.M. conceived the project. X.M. and A.K. prepared the supports. X.M. prepared and tested the membranes. X.M. performed permeation testing, XRD, and SEM. P.K. performed STEM and EDX imaging and analyzed the data with input from M.T. and K.A.M. N.M. performed the process-scale assessment with input from M.T. and P.D. M.T. directed all aspects of the project. M.T., X.M., P.K., and N.M. wrote the manuscript with input from all coauthors. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are available in the manuscript or the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6406/1008/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S10
Tables S1 and S2
References (*33*–*73*)

25 February 2018; accepted 6 July 2018
10.1126/science.aat4123

SURFACE CHEMISTRY

Regulating the femtosecond excited-state lifetime of a single molecule

K. R. Rusimova[1], R. M. Purkiss[1], R. Howes[1], F. Lee[1], S. Crampin[1,2], P. A. Sloan[1,2]*

The key to controlling reactions of molecules induced with the current of a scanning tunneling microscope (STM) tip is the ultrashort intermediate excited ionic state. The initial condition of the excited state is set by the energy and position of the injected current; thereafter, its dynamics determines the reaction outcome. We show that a STM can directly and controllably influence the excited-state dynamics. For the STM-induced desorption of toluene molecules from the Si(111)-7x7 surface, as the tip approaches the molecule, the probability of manipulation drops by two orders of magnitude. A two-channel quenching of the excited state is proposed, consisting of an invariant surface channel and a tip height–dependent channel. We conclude that picometer tip proximity regulates the lifetime of the excited state from 10 femtoseconds to less than 0.1 femtoseconds.

Using the tip of a scanning tunneling microscope (STM) to initiate chemical reactions offers a route to controllable single-molecule chemistry (*1*, *2*). Through the mechanical interaction between tip and target molecule, or by the electric field in the gap, the STM can induce molecular change across a ground-state potential energy landscape (*3*). The STM tunneling current, however, can generate excited states of a molecule and hence give enhanced specificity, and more varied outcomes, to the manipulation action [e.g., bond dissociation (*4*, *5*), isomerization (*6*), or tautomerization (*7*)]. The specificity arises by controlling the energy (*5*) or position (*7*, *8*) of the single electron (or hole) excitation within a single molecule. The ensuing molecular dynamics, and hence the final outcome, evolve naturally from that point. Having the ability to control and influence the dynamics of the excited state itself would open new paths to control matter, and its reactions, at the molecular level.

Here, we found that the lifetime of the positive ion of single toluene molecules on the Si(111)-7×7 surface can be directly controlled by the STM. By bringing the tip close to the molecule (600 to 800 pm), we regulated the excited state–mediated reaction outcome (molecular desorption) by more than two orders of magnitude. We correlate this to a reduction of the excited-state lifetime by approximately two orders of magnitude. We propose that a new electronic state is generated by the tip-molecule interaction that provides an additional decay channel for the excited state, thus quenching the excited state before its natural surface-limited lifetime elapses. We anticipate this work to be a starting point for other more complex molecular systems that yield multiple excited-state outcomes where this technique could be used to instigate, probe, and control them. The quenching process relies on fundamental quantum processes and should be applicable to a wide class of molecule/surface systems.

Multiple molecular adsorbates have been shown to react to the STM tunneling current (*1*), especially benzene and derivatives (*5*, *9*). Broadly, the probability per electron of inducing a molecular reaction is higher on the Si(111)-7×7 surface than on the Si(100)-2×1 surface and is orders of magnitude higher than on metal surfaces (*7*–*9*). On metals, lifetimes of molecular ion states are as low as 0.1 fs (*10*), but the reduced density of states in semiconductors leads to longer excited-state lifetimes. The theory of dynamics induced by electronic transition (DIET) links greater lifetimes of excited states to higher probabilities of reaction (such as bond breaking or desorption) (*11*). Benzene, chlorobenzene, and toluene on Si(111)-7×7 have all been extensively studied (*5*, *9*) and are highly sensitive to tunneling current.

Figure 1 shows a series of STM images charting the positive ion resonance (or negative-bias "hole")–induced manipulation of a single chemisorbed toluene molecule on the Si(111)-7×7 surface. At the imaging conditions used (+1 V, 100 pA), chemisorbed toluene molecules were unperturbed by the STM (*12*) and appeared as dark features against the bright spots that were the adatoms of the silicon surface. Chemisorbed toluene molecules formed a 2,5-di-σ bonding configuration with the surface, forming one covalent bond to a silicon adatom (colored red) and one to a neighboring silicon rest atom (second-layer atoms with dangling bonds) (Fig. 1A). To manipulate the molecule, during a raster scan from bottom to top, we halted the tip atop the molecule and performed current injection (–1.3 V, 900 pA) for 8 s.

Figure 1D shows the tip height variation during this process. In step i, the tip was halted above the molecule. The feedback loop was disabled, and the tip retracted by 1 nm before the voltage was ramped to the desired manipulation value with the set-point current at 20 pA. In step ii, the feedback loop was then reengaged. In step iii, the set-point current changed to the required injection current, resulting in the tip approaching the surface closer than its initial value by an amount Δz_i. Charge injection continued, and in this particular case, after 0.35 s of charge injection, the molecule-adatom bond was broken, leading to desorption. The underlying (bright) silicon adatom was exposed, causing the tip to withdraw by Δz_M to restore the set-point current (step iv). Resuming the interrupted image scan of Fig. 1B resulted in a "half-moon" feature at the molecular adsorption site, typical of a manipulation event occurring mid-scan. Subsequent image scans (Fig. 1C) had the conventional Si(111)-7×7 surface appearance, including the silicon dangling bond at the original location of the toluene molecule.

From the fraction of ~120 toluene molecules that were manipulated after an injection time t, we deduced a time-dependent probability of manipulation $P(t)$ for a single molecule, consistent with the first-order rate equation $dP(t)/dt = k[1 - P(t)]$, where k is the rate of manipulation. Figure 2A illustrates this for injection parameters of +1.6 V, 450 pA. Figure 2B illustrates how the manipulation rate k varied with tunneling current for electron injection at +1.6 V. Figure 2C presents data for hole injections at –1.3 V and at –1.0 V. For electron injection, the rate increased linearly with injection current (see fit to Fig. 2B). For hole injection at low current (2 to 10 pA), we again found a linear dependence, but beyond 10 pA the rate of manipulation was approximately constant; the fitting function of Fig. 2C is discussed below.

The number n of electrons (or holes) that drive a single-molecule manipulation (*4*) leads to a power-law dependence of the rate k with current I, $k \propto I^n$. Hence, for electron injection, the near-linear dependence $n = 0.8 \pm 0.1$ indicates a one-electron process (*9*). Similarly, at low currents, a one-hole process is responsible for desorption. However, for hole injection at currents above 10 pA, the near-constant rate implies a largely current-independent process. If the current is not driving the manipulation, what does?

The manipulation rates observed were a factor of 10^4 greater than those occurring in purely thermally driven desorption of toluene from Si(111)-7×7 (*12*). Hence, the presence of the STM tip is required for this manipulation to take place. Possible tip-molecule interactions that might drive manipulation are mechanical (i.e., a short-range chemical interaction between tip and molecule) or result from the electric field of the tunnel junction. However, we can rule out both. Figure 3A shows the tip height z during the electron and hole injections performed at

[1]Department of Physics, University of Bath, Bath BA2 7AY, UK. [2]Centre for Nanoscience and Nanotechnology, University of Bath, Bath BA2 7AY, UK.
*Corresponding author. Email: p.sloan@bath.ac.uk

different currents. The tip height z is the distance from the center of the bonding Si adatom to the center of the leading atom of the STM tip (see supplementary materials). In all cases, z exceeds 600 pm. To identify possible mechanical manipulation, we modified the manipulation experiments by setting the bias during step ii to 0 V, disabling the feedback loop, and setting z to a specific value (Fig. 3B, schematic). For each z value, we then measured the outcome of ~90 single-molecule manipulation experiments with an 8-s "exposure" of each target molecule. Little or no desorption was observed for z at or above 600 pm (Fig. 3B, shaded portion). Thus, in the height regime of the current-manipulation experiments, no mechanical manipulation occurred, and the desorption that did occur in Fig. 3B was consistent with that expected for a thermally driven process (see supplementary materials for z < 600 pm discussion).

We eliminated the possibility of an electric field–induced manipulation mechanism by modifying step ii so that, with feedback disabled, the tip retracted an additional distance from the surface. We applied a –10 V bias to generate an electric field $E \approx V/z$ in the junction comparable to that in the current-injection experiments, and whose magnitudes are shown in Fig. 3C. In this case, however, there was no current. As shown by the data in Fig. 3D, without the current, there was little or no manipulation.

A similar linear to constant rate crossover appears in two previous studies (*13*, *14*). There, tip-induced band bending (TIBB) was put forward as a possible explanation. Since then, detailed theoretical work and scanning tunneling spectroscopy show that TIBB only occurs if the semiconductor is in depletion (*15*, *16*). For our work with n-type Si, this would be for electron injection. Therefore, TIBB cannot explain our hole injection results, nor the results of (*13*). The doping level here and in (*14*) also precludes any measurable TIBB even if it occurs in the depletion regime (*17*). Instead, the model proposed here is consistent with all three reports.

The final outcome of the molecular manipulation can be either that the molecule completely leaves the surface (desorption) or that it reattaches to the surface elsewhere (diffusion). We label an outcome as diffusion if, in an "after" STM image (e.g., Fig. 1C), the manipulated molecule appeared at an adjacent binding site. All other manipulation outcomes are classified as desorption. For all injection currents used, we found a branching ratio B of the probability of desorption to diffusion that was constant throughout the hole-injection experiments, B_h = 0.037 ± 0.004. It was also constant for electron injections with B_e = 0.24 ± 0.03 over the reported range of currents. Furthermore, there was no evidence of other forms of manipulation, such as intramolecular bond dissociation (*5*), in either current regime.

Recasting the rate of manipulation in terms of the probability per injected charge of manipulation (electron or hole), $P_e = ke/I$ (where e is the magnitude of the electron charge), yields P_e as a function of the tip height z during the current injections. For electron injection, as expected for a one-electron process (Fig. 4A), P_e was fairly constant over the range of z studied. For –1.3 V hole injection (Fig. 4B), P_e exponentially increased with z (i.e., decreasing current) until at ~800 pm, we found a near-constant region. Figure 4C shows data for –1.0 V hole injections (10 to 900 pA), where for all injections, we found a similar exponential increase in the manipulation probability as the tip withdrew from the surface.

The desorption of toluene, via a DIET process, occurs in three steps (*18*): (i) excitation by capture of the injected charge by the toluene molecule; (ii) dynamics, the evolution of the ionic molecule on its excited-state potential; and (iii) detachment, with decay of the state (neutralization) leaving a vibrationally excited neutral molecule and leading to molecule-surface bond breaking and the final outcome of desorption or diffusion. Given that for the same z change, P_e was constant for electron injection, we conclude that step i was also constant for hole injection. That is, the change in the "spot size" of the tunneling current caused by the change in the value of z did not change the fraction of the current captured by the molecule. Given the invariance of the branching ratio B_h, we further conclude that step iii was the same for all the experiments shown in Fig. 4B. Within a DIET model, what remains to influence the probability of manipulation is step ii, specifically the lifetime of the excited state (*19*).

For the similar system of benzene on Si(100), Alavi *et al.* (*20*) identified a hole excited-state lifetime of ~10 fs and a probability per hole injection similar to our findings. Further, in line with theoretical predictions for long-lived excited

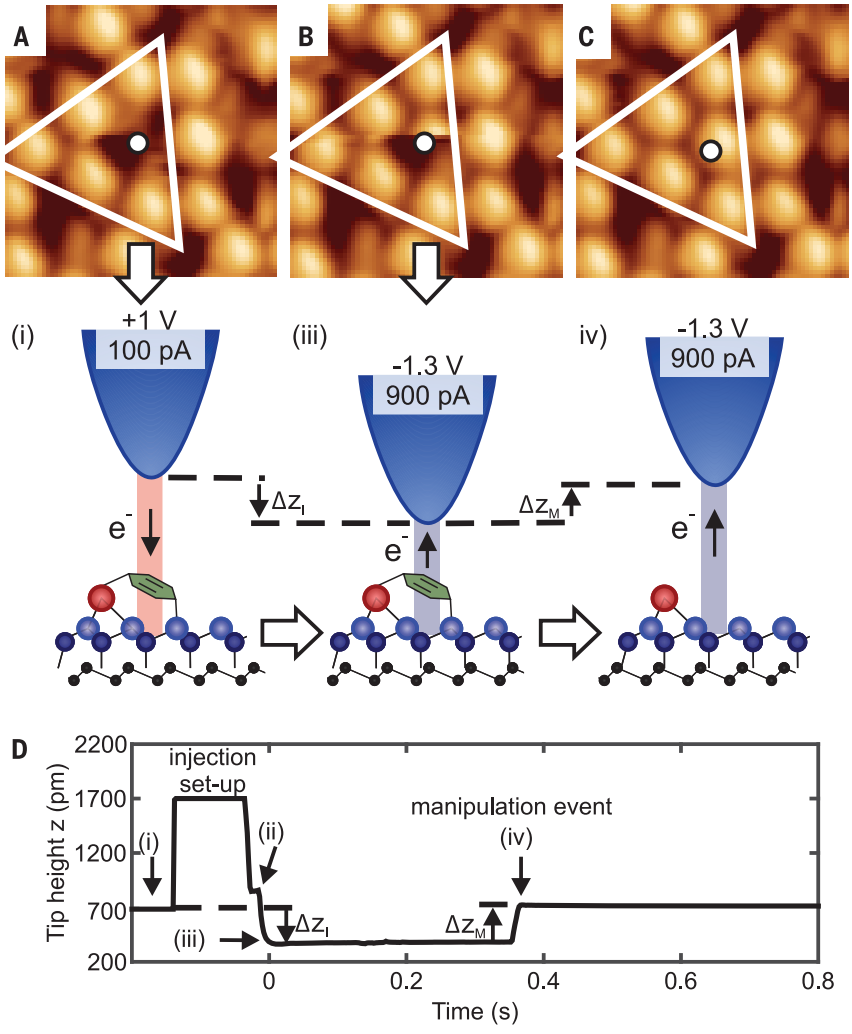


Fig. 1. STM imaging and time trace of single-molecule manipulation. (**A** to **C**) High-resolution STM images and corresponding schematic diagrams of the manipulation procedure (imaging parameters: +1 V, 100 pA, 3 nm × 3 nm). (A) Before manipulation. A half unit cell of Si(111)-7×7 is outlined; the white circle atop the missing adatom–like dark spot location indicates the position of a single toluene molecule. (B) During manipulation. (C) After manipulation. (**D**) Time trace of the tip height during charge injection.

states (*19*), Alavi *et al.* reported a monotonic and near-linear dependence of the manipulation probability on the hole excited-state lifetime. Therefore, for our hole injections, we relate our maximum (i.e., constant region) manipulation probability per hole of 320 (±10) × 10^{-9}, with an excited-state lifetime of 10 fs, and use the linear dependence $P_e = \beta\tau$ (where $\beta = 32 \times 10^{-9}$ fs^{-1}) to map our measured probability of manipulation to an excited-state lifetime τ. The result is an excited-state lifetime that changes by two orders of magnitude, from 10 fs to <0.1 fs (Fig. 4, B and C, right-side *y* axes). A value on the order of 0.1 fs is more typical of that of adsorbates on metal surfaces (*10*), indicating that the proximity of the tip transforms the molecule-semiconductor system into a metal-molecule-semiconductor system.

Studies of cyclohexadiene on Si(100) (*21*, *22*) have shown the creation of an interface electronic state at the location of the molecule as an STM tip approaches. The new state lies near the Fermi level, and in tandem with its creation, the highest occupied molecular orbital (HOMO) at −1.5 V broadens and decreases in intensity as the tip approaches closer. Given that our system also contains a π-bonding orbital on a six-member carbon ring that is di–σ-bonded to a Si substrate, we propose that at our higher currents (closest approach) a similar interfacial electronic state results in the reduced probability per hole of manipulation by providing a new decay channel for the excited state, which reduces its lifetime and concomitantly the probability of manipulation.

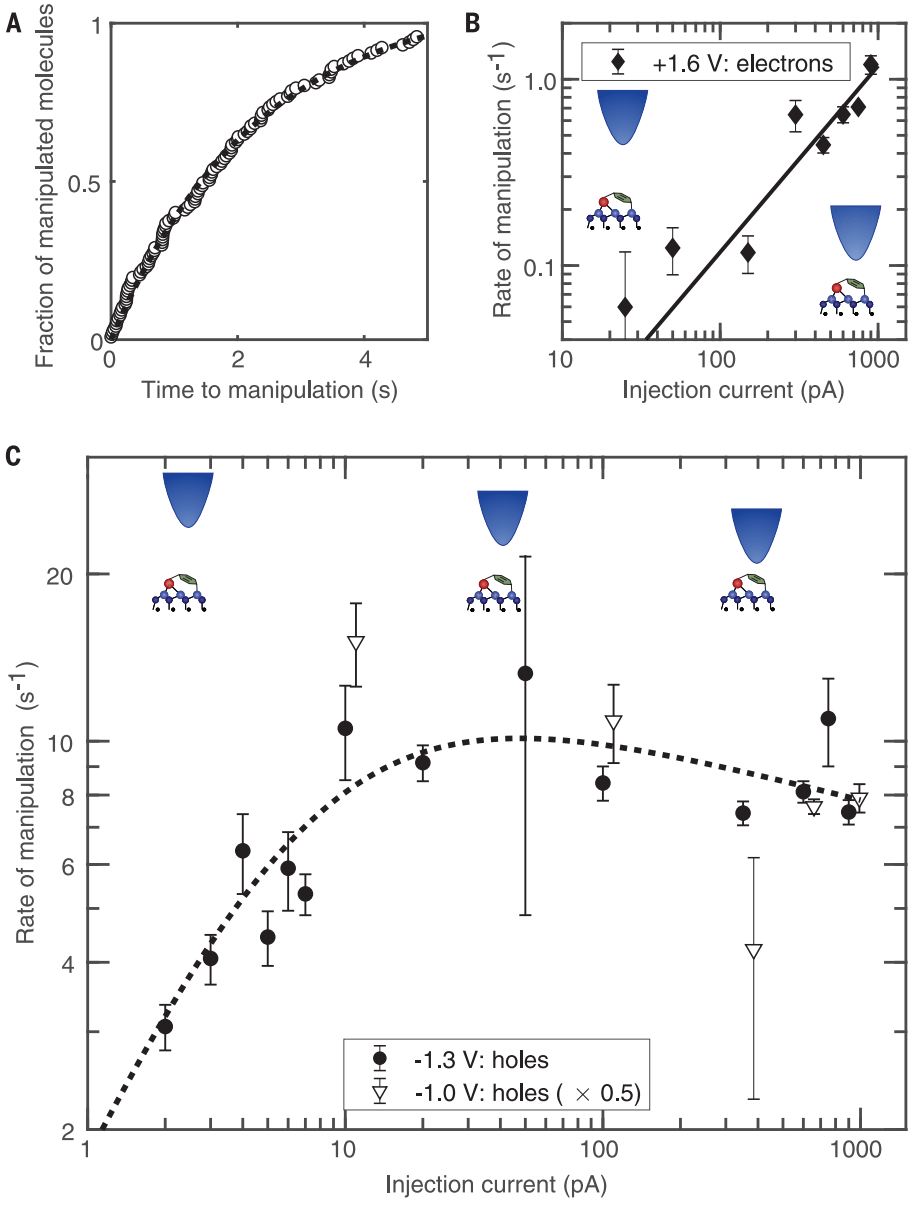


Fig. 2. Rate of manipulation. (**A**) Time dependence of the fractional manipulated molecule population (injection parameters: +1.6 V and 450 pA; 117 molecules). The dashed line shows the fit to $P(t) = 1 - \exp(-kt)$. (**B**) Rate of manipulation for electron injection at +1.6 V with a linear fit. (**C**) Rate of manipulation for hole injection at −1.3 V (solid circles) and at −1.0 V (open triangles). See text for fit details of tip-dependent model (dashed line). Error bars indicate SD.

The lifetime of an excited state is the inverse of its relaxation rate $R = 1/\tau$. We propose two components for the relaxation of the positive ion state: (i) a fixed rate arising from the presence of the surface, $R_S = 1/\tau_S$, with τ_S = 10 fs; and (ii) a *z*-dependent rate accounting for the effect of the tip, $R_T(z) = 1/\tau_T(z)$, giving $R = R_S + R_T(z)$. This tip-mediated relaxation channel will be related to the density of states of the interface state, ρ_i, through Fermi's Golden Rule. An analogous scheme is used to describe the STM excitation, direct measurement, and *z*-dependent quenching of the millisecond spin excitation in single atoms (*23*).

Figure 4D presents schematic energy level diagrams for three regimes of tip height: (I) large tip-molecule separation with surface-dominated excited-state relaxation and thus reaction, (II) intermediate-*z* range with onset of (assumed near the Fermi level) tip-dependent interface state quenching, and (III) small-*z* separation with tip-dependent interface state quenching dominating the excited-state relaxation. These three regimes are indicated in the rate dependencies of Fig. 4, A to C.

For a tip-molecule system with localized electronic structure, the force between tip and molecule has been calculated as $F \propto I^m$ with m between 1 and 2 (*24*). This calculation invoked a wave function overlap argument and should be broadly similar to the perturbative physics of the initial generation of an interface state by our STM tip. Thus, we make the connection $R_T(z) \propto \rho_i(z) \propto \exp(-2\kappa z)^m$, leading to a *z*-dependence of P_e of

$$P_e(z) = \frac{\beta\tau_S}{1 + \exp[-2\kappa m(z - z_0)]} \qquad (1)$$

where $\tau_T(z_0) = \tau_S$ and κ = 1.17 ± 0.06 A^{-1}, as found from Fig. 3A. A surface-limited model, $P_e = \beta\tau_S$, has a constant lifetime and therefore a constant manipulation probability. This surface-only model fits the +1.6 V electron injection in Fig. 4A. Below ~800 pm, there is a possible slight decrease in P_e, which suggests that the negative-ion state is also perturbed by the interface state. For −1.0 V hole injection shown in Fig. 4C, the fit is purely exponential, $P_e(z) \propto \exp(2\kappa z)$, corresponding to a tip-dominated dynamics. For −1.0 V, at all currents, the tip was near the molecule, hence the excited-state dynamics were always tip-limited. At −1.3 V, the tip was slightly farther removed from the surface. Thus, Fig. 4B shows a fit to Eq. 1 with m = 1.1 ± 0.1, and demonstrates a crossover at z_0 = 830 ± 20 pm from a tip-limited to a surface-limited regime.

Our initial finding of a near-invariant rate of manipulation can therefore be reconciled with a one-hole process. For a one-hole process, the rate is defined as $k = P_e I/e$. Combining this with the tip-dependent manipulation probability P_e of

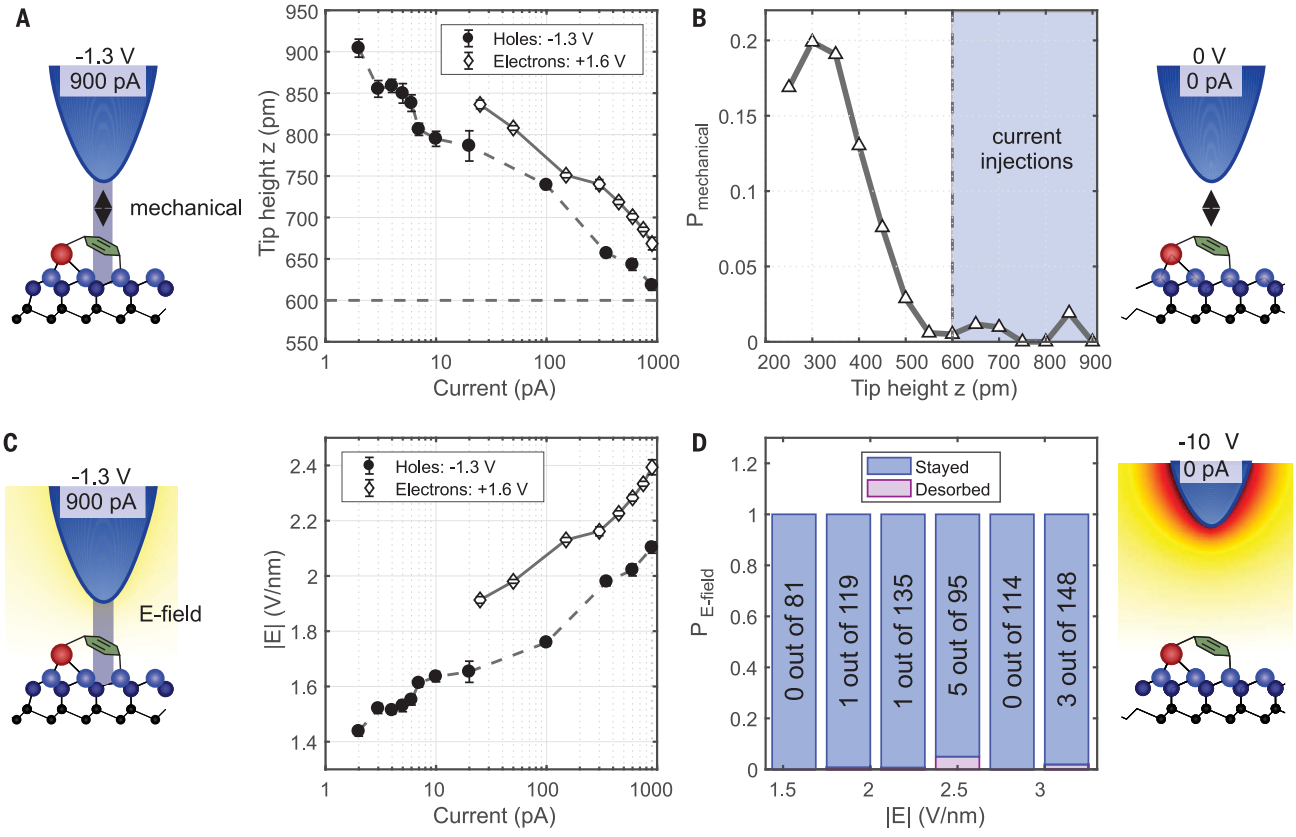


Fig. 3. Mechanical and electric-field tip-induced interactions. (**A**) Tip-surface separation as a function of tunneling current during charge injection. Solid circles, hole injection at –1.3 V; open diamonds, electron injection at +1.6 V. (**B**) Probability of manipulation after 8 s in the mechanical presence of the tip (0 V, 0 pA). (**C**) Estimated electric field (magnitude) in the junction between tip and surface as a function of the current *I* during the charge injection manipulation experiments. Solid circles, hole injection at –1.3 V; open diamonds, electron injection at +1.6 V. (**D**) Probability of manipulation after 8 s with only the electric field interaction (–10 V, 0 pA). Error bars indicate SD; some error bars are too small to see.

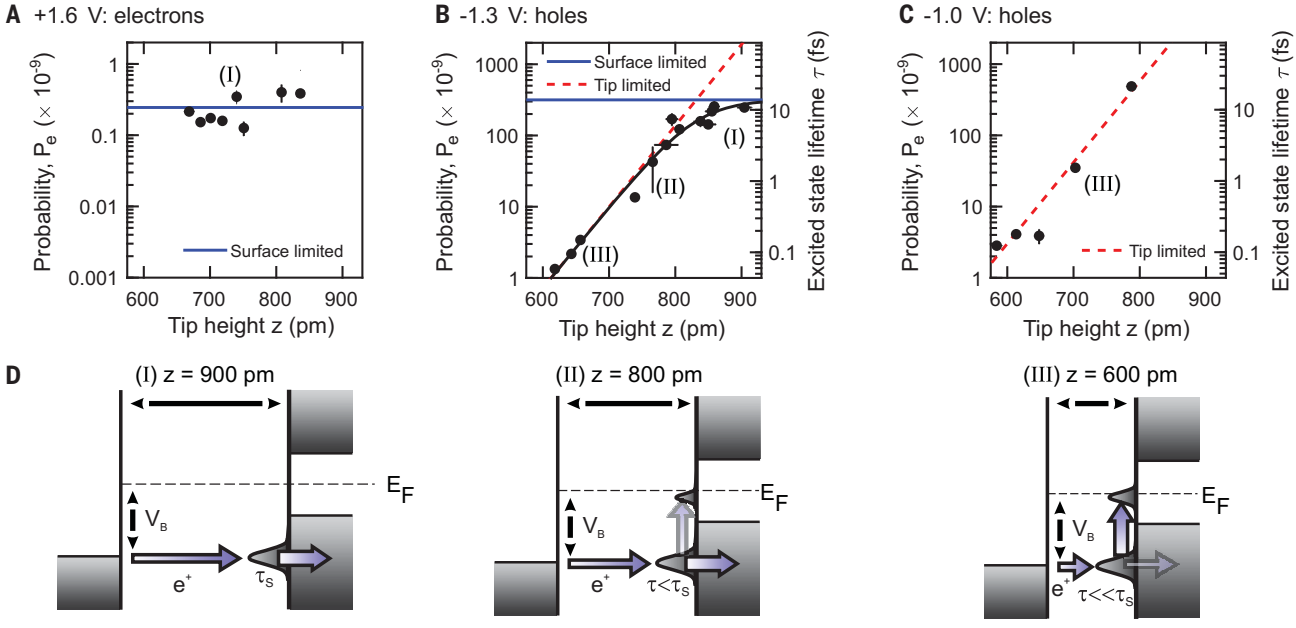


Fig. 4. Manipulation suppression at close tip proximity. (**A**) +1.6 V electron injection data from Fig. 2B recast as probability per electron as a function of the tip height. (**B**) –1.3 V hole injection data from Fig. 2C, recast as in (A). (**C**) –1.0 V hole injection. Blue lines in (A) and (B) show surface-limited excited-state dynamics; dashed red lines in (B) and (C) model tip-limited dynamics. Black curve in (B) is a fit to Eq. 1. Right-side *y* axes of (B) and (C) show the inferred excited-state lifetime τ of the positive ion with a value of 10 fs for purely surface-limited dynamics. (**D**) Schematic energy level diagrams depicting three regimes of tip manipulation suppression. V_B is the bias voltage applied to the tip relative to the sample Fermi level E_F.

Eq. 1 gives the fit to the rate of manipulation (dashed line) in Fig. 2C. At large tip-molecule separation (low current), the traditional linear $k \propto I$ dependence is evident. For higher currents that led to a closer tip and quenching of the molecular excited state, the rate of manipulation became $k = I/I^m = I^{-0.1\pm0.1}$, giving the largely rate-invariant region of manipulation (Fig. 2C).

Molecules that require more than one electron (or hole) for manipulation, such as C-Cl dissociation of chlorobenzene (*5*) or diffusion of NH_3 (*25*), should naturally be more sensitive to any tip modification of the excited state. Alongside other semiconducting molecule/surface systems, we would also expect any molecule/surface system that displays a long-lived excited state—for example, molecule/single–atomic layer insulator/metal systems (*26*)—to be sensitive to tip-induced modification of the excited state.

REFERENCES AND NOTES

1. A. J. Mayne, G. Dujardin, G. Comtet, D. Riedel, *Chem. Rev.* **106**, 4355–4378 (2006).
2. Y. Jiang, Q. Huan, L. Fabris, G. C. Bazan, W. Ho, *Nat. Chem.* **5**, 36–41 (2013).
3. R. Garcia, A. W. Knoll, E. Riedo, *Nat. Nanotechnol.* **9**, 577–587 (2014).
4. B. C. Stipe *et al*., *Phys. Rev. Lett.* **78**, 4410–4413 (1997).
5. P. A. Sloan, R. E. Palmer, *Nature* **434**, 367–371 (2005).
6. V. Simic-Milosevic, M. Mehlhorn, K. H. Rieder, J. Meyer, K. Morgenstern, *Phys. Rev. Lett.* **98**, 116102 (2007).
7. P. Liljeroth, J. Repp, G. Meyer, *Science* **317**, 1203–1206 (2007).
8. M. Lastapis *et al*., *Science* **308**, 1000–1003 (2005).
9. K. R. Rusimova, P. A. Sloan, *Nanotechnology* **28**, 054002 (2017).
10. L. Bartels *et al*., *Phys. Rev. Lett.* **80**, 2004–2007 (1998).
11. P. Saalfrank, *Chem. Rev.* **106**, 4116–4159 (2006).
12. D. Lock, S. Sakulsermsuk, R. E. Palmer, P. A. Sloan, *J. Phys. Condens. Matter* **27**, 054003 (2015).
13. B. C. Stipe, M. A. Rezaei, W. Ho, *Phys. Rev. Lett.* **79**, 4397–4400 (1997).
14. X. Lu, J. C. Polanyi, J. S. Yang, *Nano Lett.* **6**, 809–814 (2006).
15. J. Mysliveček *et al*., *Phys. Rev. B* **73**, 161302 (2006).
16. R. M. Feenstra, S. Gaan, G. Meyer, K. H. Rieder, *Phys. Rev. B* **71**, 125316 (2005).
17. R. M. Feenstra, *Phys. Rev. B* **50**, 4561–4570 (1994).
18. R. E. Palmer, P. J. Rous, *Rev. Mod. Phys.* **64**, 383–440 (1992).
19. P. Saalfrank, G. Boendgen, C. Corriol, T. Nakajima, *Faraday Discuss.* **117**, 65–83 (2000).
20. S. Alavi *et al*., *Phys. Rev. Lett.* **85**, 5372–5375 (2000).
21. B. Naydenov, L. C. Teague, P. Ryan, J. J. Boland, *Nano Lett.* **6**, 1752–1756 (2006).
22. P. M. Ryan, L. C. Teague, B. Naydenov, D. Borland, J. J. Boland, *Phys. Rev. Lett.* **101**, 096801 (2008).
23. W. Paul *et al*., *Nat. Phys.* **13**, 403–407 (2017).
24. P. Jelínek, M. Ondráček, F. Flores, *J. Phys. Condens. Matter* **24**, 084001 (2012).
25. J. I. Pascual, N. Lorente, Z. Song, H. Conrad, H. P. Rust, *Nature* **423**, 525–528 (2003).
26. J. Repp, G. Meyer, S. Paavilainen, F. E. Olsson, M. Persson, *Science* **312**, 1196–1199 (2006).

ACKNOWLEDGMENTS

We thank R. Palmer, D. Bird, and D. Wolverson for fruitful discussions. **Funding:** Supported by EPSRC grant EP/K00137X/1 (P.A.S.), a University of Bath studentship (K.R.R.), and EPSRC CDT CMP grant EP/L015544/1 (R.M.P.). **Author contributions:** K.R.R. was the primary experimentalist and performed the analysis. R.M.P., R.H., and F.L. performed subsets of the experiments; S.C. provided theoretic support and data interpretation; P.A.S. led the team, designed the experiment and the analysis; and K.R.R., S.C., and P.A.S. wrote the manuscript. **Competing interests:** Authors declare no competing interests. **Data and materials availability:** All data supporting this study are openly available from the University of Bath data archive at https://doi.org/10.15125/BATH-00527.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6406/1012/suppl/DC1
Materials and Methods
Supplementary Text
References (*27*–*37*)

30 April 2018; accepted 9 July 2018
10.1126/science.aat9688

GALAXY FEEDBACK

Fast molecular outflow from a dusty star-forming galaxy in the early Universe

J. S. Spilker[1,2]*, M. Aravena[3], M. Béthermin[4], S. C. Chapman[5], C.-C. Chen[6], D. J. M. Cunningham[5,7], C. De Breuck[6], C. Dong[8], A. H. Gonzalez[8], C. C. Hayward[9,10], Y. D. Hezaveh[11], K. C. Litke[2], J. Ma[12], M. Malkan[13], D. P. Marrone[2], T. B. Miller[5,14], W. R. Morningstar[11], D. Narayanan[8], K. A. Phadke[15], J. Sreevani[15], A. A. Stark[10], J. D. Vieira[15], A. Weiß[16]

Galaxies grow inefficiently, with only a small percentage of the available gas converted into stars each free-fall time. Feedback processes, such as outflowing winds driven by radiation pressure, supernovae, or supermassive black hole accretion, can act to halt star formation if they heat or expel the gas supply. We report a molecular outflow launched from a dust-rich star-forming galaxy at redshift 5.3, 1 billion years after the Big Bang. The outflow reaches velocities up to 800 kilometers per second relative to the galaxy, is resolved into multiple clumps, and carries mass at a rate within a factor of 2 of the star formation rate. Our results show that molecular outflows can remove a large fraction of the gas available for star formation from galaxies at high redshift.

The formation of realistic populations of galaxies requires one or more forms of self-regulating feedback to suppress the conversion of gas into stars. Cosmological simulations invoke various mechanisms to regulate star formation, in the form of energy deposition and wind launching linked to supermassive black hole activity, supernovae, and/or radiation pressure from massive stars. The strength and scalings of these processes play critical roles in the evolution of galaxies by regulating the growth of stellar mass relative to the dark-matter halo, connecting the properties of central black holes to their host galaxies, and enriching the circumgalactic medium with heavy elements (*1*–*4*).

Outflowing winds of gas are ubiquitous in nearby galaxies. The gas in outflows spans many orders of magnitude in temperature and density (*5*–*7*), and different components of the winds are observable from x-ray to radio wavelengths. Observing winds in the distant Universe is difficult: Not only are the spectral features faint, but outflow tracers observed in emission may be less reliable because of the ongoing processes of galaxy assembly (*8*). The constrained geometry of absorption lines provides signatures of inflowing and outflowing material, but these lines have thus far eluded detection.

Passive galaxies with stellar masses of ~10^{11} $M_\odot$ (where $M_\odot$ is the solar mass), low star formation rates (SFRs) ($\lesssim 10 M_\odot$ yr^{-1}), and stellar ages of ~0.8 billion years were already in place when the Universe was 2 billion years old, implying that these galaxies formed stars at rates of hundreds of $M_\odot$ per year before $z = 5$ (where z is the redshift) (*9*). Galaxies with such high SFRs are extremely rare in rest-ultraviolet surveys, implying that such galaxies are incapable of becoming sufficiently massive by $z \sim 4$ to reproduce the observed passive population. This suggests a connection between early passive galaxies and high-redshift dusty star-forming galaxies (DSFGs) observed at far-infrared wavelengths (*10*–*12*). With a redshift distribution that includes a substantial number of objects at $z > 4$ (*13*, *14*), DSFGs represent a plausible progenitor population of the earliest passive galaxies. If this evolutionary connection is correct, many DSFGs should show signs of the feedback process(es) acting to suppress their rapid star formation.

We present observational evidence of a massive molecular wind being launched from SPT–S J231921–5557.9 (hereinafter referred to as SPT2319–55), a DSFG observed when the Universe was only 1 billion years old. SPT2319–55 was discovered in the 2500-square-degree South Pole Telescope survey (*15*) on the basis of its thermal dust emission. Earlier observations of this source from the Atacama Large Millimeter/submillimeter Array (ALMA) determined its redshift to be $z_{\rm source} = 5.293$ (*14*) and showed that it is gravitationally lensed by an intervening foreground galaxy (*16*). As is typical for these objects, SPT2319–55 is gas rich, containing ~1.2×10^{10} $M_\odot$ of molecular gas and ~1.2×10^{8} $M_\odot$ of dust, and is forming stars very rapidly, with an SFR of ~790 $M_\odot$ yr^{-1} (*17*) (these values account for the lensing magnification).

We used ALMA to observe the rest-frame 119-μm ground-state doublet transition of the hydroxyl molecule, OH, and the thermal dust emission at this wavelength (*17*). This transition is a good tracer of gas flows in nearby galaxies (*18*, *19*). The ALMA observations reach a spatial resolution of 0.25 by 0.4 arc sec and resolve the lensed images of SPT2319–55 (Fig. 1). We detected a molecular outflow from SPT2319–55, seen in blueshifted absorption against the bright dust continuum emission, a signature of outflowing molecular material (Fig. 1).

We fit the spectrum in Fig. 1 with the sum of two velocity components, each consisting of two equal-amplitude Gaussian profiles separated by 520 km s^{-1}, the separation of the two components of the OH doublet (*17*). As is common practice for low-redshift observations of OH, we assigned the higher-frequency component of the doublet to the systemic velocity of the galaxy. Because the redshift of this source is known to better than 50 km s^{-1} from other observations (*14*), we fixed the velocity offset of one pair of Gaussian profiles to the systemic velocity. This component represents absorption due to gas from within the galaxy, with a fitted full-width-at-half-maximum (FWHM) linewidth of 330 ± 80 km s^{-1}. We allowed a velocity offset for the second pair of Gaussian profiles; this component represents the blueshifted molecular outflow. We found that this second component is blueshifted relative to the galaxy by 440 ± 50 km s^{-1} and derived a maximum velocity of ~800 km s^{-1} (*17*).

The ALMA observations spatially resolve and clearly detect the dust continuum even in relatively narrow velocity channels. Because SPT2319–55 is gravitationally lensed by a foreground galaxy, we determine its intrinsic structure by using a lens modeling technique that represents the galaxy as an array of pixels (*17*, *20*). In addition to the line-free continuum emission, we also reconstructed the OH absorption components by using velocity ranges relative to the higher-frequency OH transition of −700 to −200 km s^{-1} for the wind component and +300 to +700 km s^{-1} for the internal component; the latter velocity range corresponds to velocities of −220 to +180 km s^{-1} relative to the lower-frequency transition and traces gas within SPT2319–55. These

[1]Department of Astronomy, University of Texas at Austin, 2515 Speedway Stop C1400, Austin, TX 78712, USA. [2]Steward Observatory, University of Arizona, 933 North Cherry Avenue Tucson, AZ 85721, USA. [3]Núcleo de Astronomía, Facultad de Ingeniería, Universidad Diego Portales, Av. Ejército 441, Santiago, Chile. [4]Aix Marseille Univ., Centre National de la Recherche Scientifique, Laboratoire d'Astrophysique de Marseille, Marseille, France. [5]Department of Physics and Atmospheric Science, Dalhousie University, Halifax, Nova Scotia, Canada. [6]European Southern Observatory, Karl Schwarzschild Straße 2, 85748 Garching, Germany. [7]Department of Astronomy and Physics, Saint Mary's University, Halifax, Nova Scotia, Canada. [8]Department of Astronomy, University of Florida, Bryant Space Sciences Center, Gainesville, FL 32611, USA. [9]Center for Computational Astrophysics, Flatiron Institute, 162 Fifth Avenue New York, NY 10010, USA. [10]Harvard-Smithsonian Center for Astrophysics, 60 Garden St., Cambridge, MA 02138, USA. [11]Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, Stanford, CA 94305, USA. [12]Department of Physics and Astronomy, University of California, Irvine, CA 92697, USA. [13]Department of Physics and Astronomy, University of California, Los Angeles, CA 90095, USA. [14]Department of Astronomy, Yale University, 52 Hillhouse Avenue New Haven, CT 06511, USA. [15]Department of Astronomy, University of Illinois, 1002 West Green Street Urbana, IL 61801, USA. [16]Max-Planck-Institut für Radioastronomie, Auf dem Hügel 69, D-53121 Bonn, Germany.
*Corresponding author. Email: spilkerj@gmail.com

velocity ranges fairly cleanly separate the wind and internal absorption (*17*).

The doubly-imaged continuum emission in Fig. 1 is consistent with the lensing of a single background galaxy, with a well-determined extent, by a single foreground lens. The reconstructions of the dust continuum and each absorption component are shown in Fig. 2. The continuum is dominated by a single bright region ~1.2 kpc in diameter (*17*), with flux density $S_{400\mathrm{GHz}}$ = 9.0 mJy (1 Jy = 10^{-26} W m^{-2} Hz^{-1}) magnified by a factor μ = 5.8. The internal absorption, arising from gas within the galaxy, is concentrated toward the center of the object. This is to be expected, as the continuum is brightest and gas column densities are highest toward the nuclear region, making the detection of absorption easier.

The geometry of the molecular outflow is more complex and is not confined to the center of the galaxy. Instead, it is clustered into multiple clumps, separated from each other by a few hundred parsecs, corresponding to ≈2.5 FWHM resolution elements (*17*). Because absorption is easier to detect against a strong continuum, we would expect a geometry similar to that of the internal absorption if the continuum were the limiting factor in the reconstruction. Tests using mock data also show that absorbing components weaker than the outflow are well recovered by our reconstructions (*17*).

The overall covering fraction of the wind is high—80% of pixels with a continuum signal-to-noise ratio of >8 have significant wind absorption (although these pixels are not all independent) (*17*). If the covering fraction along the line of sight we observe is the same as that for the rest of the source, this implies a total wind opening angle of ~0.8 × 4π sr. On the other hand, we can set a rough lower limit on the opening angle if we assume that we have observed the entirety of the outflow and that the molecular material is destroyed by the time it reaches a few galaxy radii away, as in local starburst galaxies (*6*). If the maximum radius of the outflowing material is 2 galaxy radii, for example, then the solid angle of a sphere of this radius subtended by the source is $(\pi r_{\mathrm{eff}}^2)/\left(4\pi(2r_{\mathrm{eff}})^2\right) \times 4\pi$ sr, with r_{eff} representing the radius of the galaxy. With an 80% covering fraction of the source along the line of sight, this implies a minimum opening angle of 0.2π sr.

In low-redshift galaxies, the 119-μm OH transitions are very optically thick (*21*), and the absorption depth thus directly corresponds to the covering fraction of the continuum. The difference between the 80% covering fraction from our lens modeling and the peak wind depth (~15%) mirrors the discrepancy between the typical absorption depths and the high detection rate of

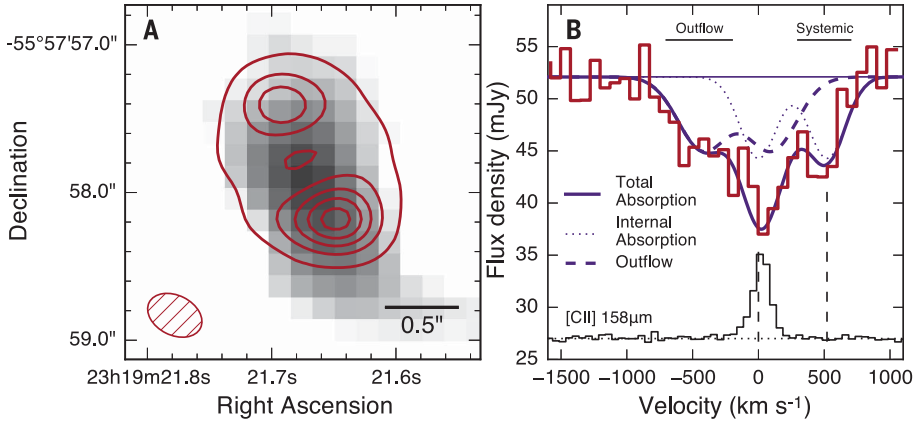


Fig. 1. ALMA 400-GHz continuum image and OH spectrum of SPT2319−55. (**A**). The ALMA continuum data (red contours) overlaid on a 2.2-μm image of the foreground lens galaxy. Contours are drawn at 10, 30, 50, 70, and 90% of the peak value; the data reach a peak signal-to-noise ratio of ~140. The synthesized beam is shown by a striped ellipse at the lower left. (**B**) The integrated apparent (not corrected for lensing magnification) OH 119-μm spectrum of SPT2319−55, with the velocity scale relative to the higher-frequency component of the OH doublet. The rest velocities of the two doublet components are shown with vertical dashed lines, and we show an ALMA [CII] spectrum of this source (*17*) as an indication of the linewidth of this galaxy due to internal gas motions (with arbitrary vertical normalization). We fit the OH 119-μm spectrum as described in the text; the navy dotted line shows the component due to gas within the galaxy, the navy dashed line the blueshifted outflow component, and the navy solid line the total absorption profile. We mark the velocity ranges for which we created lens models with horizontal bars above the continuum.

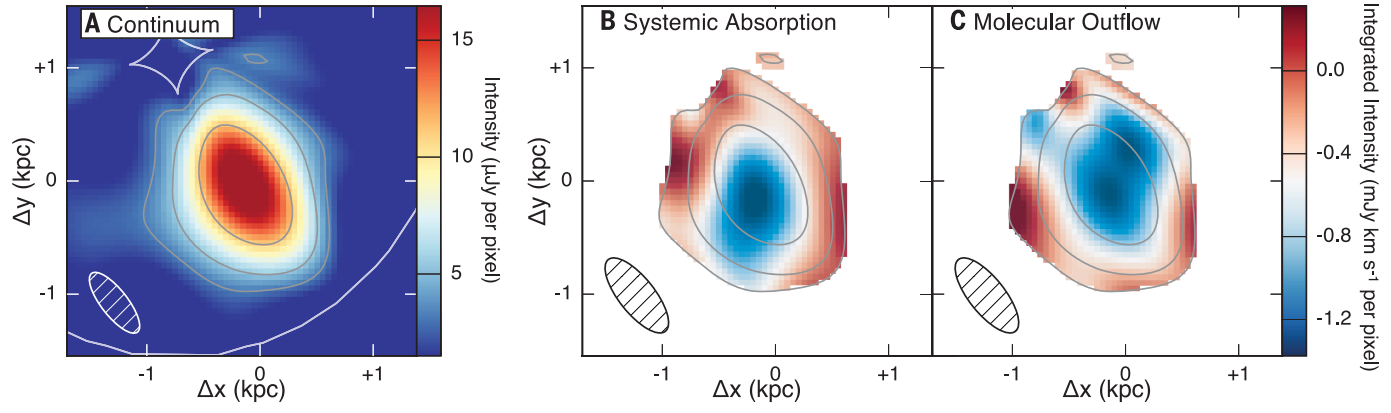


Fig. 2. Lensing reconstruction of the continuum and OH absorption in SPT2319−55. (**A**) Source-plane structure of the rest-frame 119-μm continuum emission after modeling of the effects of gravitational lensing. Contours are shown at signal-to-noise ratios of 8, 16, and 32; these contours are repeated in subsequent panels. The ellipse at the lower left shows the effective resolution of the reconstruction (*17*). The lensing caustics (lines of theoretically infinite magnification) are also shown. Coordinates are offset relative to the ALMA phase center. (**B**) Reconstruction of the absorption due to gas internal to SPT2319−55. (**C**) Reconstruction of the molecular outflow. Because the reconstruction of the absorption requires the presence of continuum emission, in panels (B) and (C) we mask pixels with a continuum signal-to-noise ratio of <8. We separate the internal and outflow absorption by using the velocity ranges labeled in Fig. 1. Ellipses at the lower left indicate the effective spatial resolution of the absorption reconstructions (*17*).

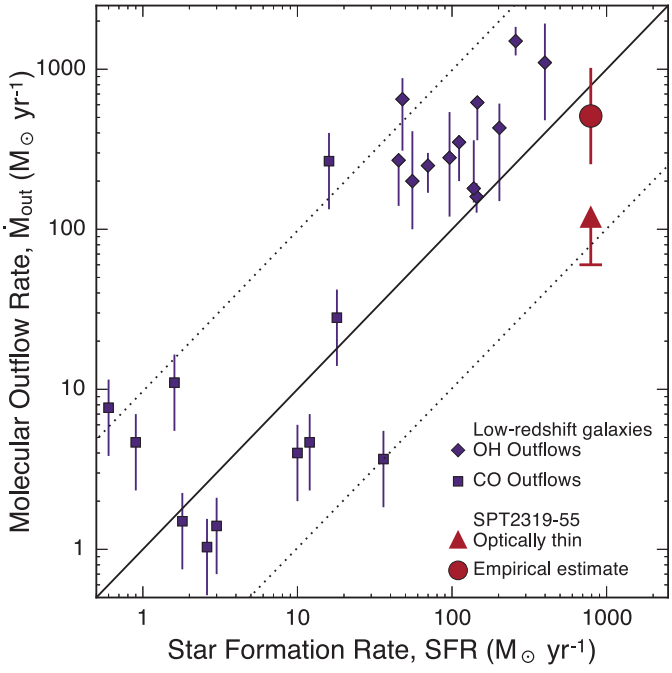


Fig. 3. Molecular outflow rates as a function of the star formation rate. Outflow rates are from both OH and CO (*21*, *23*). The solid line shows the one-to-one relation, with dotted lines representing a factor of 10 above and below. The low-redshift samples include both sources dominated by star formation and sources with active galactic nuclei. Error bars show estimated standard uncertainties on the measurements.

winds in low-redshift objects studied in OH (*19*, *22*). If the OH absorption in SPT2319–55 is also optically thick, we expect that the wind contains substantial substructure on scales below our effective resolution limit, with most of the absorption arising from small, highly optically thick clumps. Alternatively, because the outflow reconstruction spans a wide range of velocities, it is possible that the covering fraction is lower in narrower velocity ranges. In this case, we also expect substantial substructure on smaller scales, as the lower covering fractions in narrow velocity bins must still sum to the high overall covering fraction we have observed. On the other hand, if the OH excitation temperature is high, the OH molecules can substantially fill in the absorption profile with re-emitted 119-μm photons, reducing the absorption strength while maintaining a high covering fraction without small-scale structure in the wind, although this implies high densities in the absorbing gas.

We estimated the mass outflow rate by using a simple model for the outflow geometry (*17*). The primary uncertainties in this estimate are the unknown optical depth of the OH 119-μm doublet and the detailed geometry of the wind. In the limiting case of optically thin absorption, we find a minimum mass outflow rate $\dot{M}_{out} \gtrsim 60\,M_\odot\,yr^{-1}$. A more likely outflow rate, on the basis of the empirical correlation between the OH equivalent width and $\dot{M}_{out}$ in low-redshift dusty galaxies (*17*), is $\dot{M}_{out} \sim 510\,M_\odot yr^{-1}$. Though highly uncertain (*17*), this value is within a factor of 2 of the SFR of SPT2319–55, indicating that the wind is capable of depleting the molecular gas reservoir on a time scale similar to that of star formation itself. We compare the SFR and molecular outflow rates we derived for SPT2319–55 under both assumptions in Fig. 3, along with molecular outflow rates for low-redshift objects (*21*, *23*).

The mass loading factor of the wind, $\dot{M}_{out}/SFR \sim 0.7$, is high, even accounting for molecular material alone; the inclusion of unobserved wind phases, namely, neutral atomic and ionized gases, would increase this value still further. High mass loading factors are expected from simulations of galaxies (*24*, *25*), which require strong feedback to prevent the overproduction of stars. On the other hand, both the origin of the molecules in the SPT2319–55 wind (whether entrained in hotter material or formed in situ) and the driving source are unclear. Current data place an upper limit of 30% on the contribution of an active galactic nucleus to the total luminosity of SPT2319–55 but cannot rule out nuclear activity at a lower level (*17*).

Despite the high overall outflow rate, it is likely that only a small fraction of the wind mass is sufficiently fast-moving to escape the gravitational potential of the galaxy. Given the molecular gas mass of SPT2319–55, assuming a typical gas fraction for DSFGs at this redshift (*26*) and using the size from the source reconstruction yield an escape velocity from the galaxy of ~650 km s^{-1}. The fraction of the absorption at speeds greater than this value indicates that ~10% of the outflowing material will escape SPT2319–55, under the same assumptions as those for the outflow as a whole. This escape fraction is within the range observed in local galaxies (*27*). The remainder of the material, therefore, will stay within the galaxy's dark-matter halo.

Our results show that self-regulating feedback is acting to disrupt and remove the molecular gas in SPT2319–55 and will likely suppress the rapid star formation in this galaxy in $\lesssim$100 million years. Whether this is sufficient to quench the star formation on a more permanent basis is less clear. A large fraction of the wind material is likely to remain within the galaxy, capable of fueling future star formation at later times. On the other hand, SPT2319–55 likely resides in a dark-matter halo of mass ~$10^{12}\,M_\odot$ (*28*), sufficiently massive to shock infalling gas to high temperatures and prevent its accretion (*29*). Other forms of feedback, such as from an active galactic nucleus, may also act to maintain the low SFR after the initial suppression (*30*). Molecular outflows like the one we have observed are likely an important step in the suppression of star formation and the emergence of early passive galaxies.

REFERENCES AND NOTES

1. N. Murray, E. Quataert, T. A. Thompson, *Astrophys. J.* **618**, 569–585 (2005).
2. P. F. Hopkins, L. Hernquist, T. J. Cox, D. Kereš, *Astrophys. J.* **175** (suppl.), 356–389 (2008).
3. J. Kormendy, L. C. Ho, *Annu. Rev. Astron. Astrophys.* **51**, 511–653 (2013).
4. D. Ceverino *et al.*, *Mon. Not. R. Astron. Soc.* **457**, 2605–2612 (2016).
5. S. Veilleux, G. Cecil, J. Bland-Hawthorn, *Annu. Rev. Astron. Astrophys.* **43**, 769–826 (2005).
6. A. K. Leroy *et al.*, *Astrophys. J.* **814**, 83 (2015).
7. E. E. Schneider, B. E. Robertson, *Astrophys. J.* **834**, 144 (2017).
8. D. Narayanan *et al.*, *Nature* **525**, 496–499 (2015).
9. C. M. S. Straatman *et al.*, *Astrophys. J.* **783**, L14 (2014).
10. A. W. Blain, S. C. Chapman, I. Smail, R. Ivison, *Astrophys. J.* **611**, 725–731 (2004).
11. L. J. Tacconi *et al.*, *Astrophys. J.* **680**, 246–262 (2008).
12. K. Glazebrook *et al.*, *Nature* **544**, 71–74 (2017).
13. J. D. Vieira *et al.*, *Nature* **495**, 344–347 (2013).
14. M. L. Strandet *et al.*, *Astrophys. J.* **822**, 80 (2016).
15. L. M. Mocanu *et al.*, *Astrophys. J.* **779**, 61 (2013).
16. J. S. Spilker *et al.*, *Astrophys. J.* **826**, 112 (2016).
17. See supplementary materials.
18. E. Sturm *et al.*, *Astrophys. J.* **733**, L16 (2011).
19. S. Veilleux *et al.*, *Astrophys. J.* **776**, 27 (2013).
20. Y. D. Hezaveh *et al.*, *Astrophys. J.* **823**, 37 (2016).
21. E. González-Alfonso *et al.*, *Astrophys. J.* **836**, 11 (2017).
22. H. W. W. Spoon *et al.*, *Astrophys. J.* **775**, 127 (2013).
23. C. Cicone *et al.*, *Astron. Astrophys.* **562**, A21 (2014).
24. A. L. Muratov *et al.*, *Mon. Not. R. Astron. Soc.* **454**, 2691–2713 (2015).
25. C. C. Hayward, P. F. Hopkins, *Mon. Not. R. Astron. Soc.* **465**, 1682–1698 (2017).
26. M. Aravena *et al.*, *Mon. Not. R. Astron. Soc.* **457**, 4406–4420 (2016).
27. K. Alatalo, *Astrophys. J.* **801**, L17 (2015).
28. P. S. Behroozi, R. H. Wechsler, C. Conroy, *Astrophys. J.* **770**, 57 (2013).
29. A. Dekel, Y. Birnboim, *Mon. Not. R. Astron. Soc.* **368**, 2–20 (2006).
30. A. C. Fabian, *Annu. Rev. Astron. Astrophys.* **50**, 455–489 (2012).

ACKNOWLEDGMENTS

We gratefully acknowledge the allocation of computer time from UA High Performance Computing at the University of Arizona, and we thank the Texas Advanced Computing Center (TACC) at the University of Texas at Austin for providing high-performance computing resources that have contributed to the results reported here. **Funding:** ALMA is a partnership of ESO (representing its member states), the NSF (USA), and NINS (Japan), together with the NRC (Canada) and NSC and ASIAA (Taiwan), in cooperation with the Republic of Chile. The Joint ALMA Observatory is operated by ESO, AUI/NRAO, and NAOJ. The South Pole Telescope is supported by the NSF through grant PLR-1248097, with partial support through PHY-1125897, the Kavli Foundation, and Gordon and Betty Moore Foundation grant GBMF 947. The Australia Telescope Compact Array is part of the Australia Telescope National Facility, which is funded by the Australian government for operation as a National Facility

managed by CSIRO. J.S.S. thanks the McDonald Observatory at the University of Texas at Austin for support through a Smith fellowship. J.S.S., K.C.L., D.P.M., J.S., and J.D.V. acknowledge support from the NSF under grant AST-1312950; K.C.L. and D.P.M. also acknowledge support under AST-1715213, and J.S. and J.D.V. also acknowledge support under AST-1716127. C.C.H. acknowledges support from the Flatiron Institute, which is supported by the Simons Foundation. Y.D.H. is a Hubble fellow.
Author contributions: J.S.S. proposed the ALMA OH observations, performed the data reduction and lensing analysis, and wrote the manuscript. D.P.M. and K.C.L. led the observations of [CII] shown in Fig. 1. M.A. led the CO observations used to calculate the molecular gas mass. K.A.P. and J.D.V. performed modeling of the source spectral energy distribution. All authors discussed the results and provided comments on the figures and text. Authors are ordered alphabetically after J.S.S. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** This paper makes use of the following ALMA data: ADS/JAO.ALMA#2016.1.00089.S and ADS/JAO.ALMA#2016.1.01499.S, archived at https://almascience.nrao.edu/alma-data/archive. Pixelated reconstructions were performed by using a lens modeling tool developed by a subset of the authors and additional non-authors. Developers of this tool include authors Y.D.H. and W.R.M. and non-authors N. Dalal, G. Holder, and P. Marshall. A binary executable version of the lensing tool is provided at https://github.com/yasharhezaveh/Ripple/releases. Reduced data products, lens tool data files, and lensing reconstructions are provided at https://github.com/jspilker/s18_outflow.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6406/1016/suppl/DC1
Materials and Methods
Figs. S1 to S8
Tables S1 and S2
References (*31*–*61*)

3 October 2017; accepted 13 July 2018
10.1126/science.aap8900

ENERGY

Climate model shows large-scale wind and solar farms in the Sahara increase rain and vegetation

Yan Li[1,2,3]*, Eugenia Kalnay[1,4]*, Safa Motesharrei[1,4,5]*, Jorge Rivas†, Fred Kucharski[6], Daniel Kirk-Davidoff[1], Eviatar Bach[1,4], Ning Zeng[1,7]

Wind and solar farms offer a major pathway to clean, renewable energies. However, these farms would significantly change land surface properties, and, if sufficiently large, the farms may lead to unintended climate consequences. In this study, we used a climate model with dynamic vegetation to show that large-scale installations of wind and solar farms covering the Sahara lead to a local temperature increase and more than a twofold precipitation increase, especially in the Sahel, through increased surface friction and reduced albedo. The resulting increase in vegetation further enhances precipitation, creating a positive albedo–precipitation–vegetation feedback that contributes ~80% of the precipitation increase for wind farms. This local enhancement is scale dependent and is particular to the Sahara, with small impacts in other deserts.

Limiting global warming to 2°C is essential for mitigating excessive damages from climate change (*1*–*3*). Major global efforts and long-term policies are needed to attain the corresponding level of decarbonization (*4*–*6*). Renewable energy sources such as wind and solar power have become viable options (*7*) because of their abundant supply and wide availability on Earth (*8*, *9*). Extracting a small fraction of the solar and wind energy available on Earth would be more than enough to meet the total global demand of energy in all forms. This opens the possibility of powering the world entirely with wind and solar energy, which is possible and has been discussed in the literature (*9*–*13*).

To substitute for the fossil fuels that currently still dominate worldwide electricity generation, as well as transportation, heating, and industrial energy demands, more large-scale wind and solar farms would need to be installed throughout the world. The installed wind turbines and photovoltaic panels would cover the land and modify land surface properties (in particular, surface roughness and albedo, respectively) and, if large enough, could have unintended consequences on local and regional climate (*14*–*16*). Previous modeling studies have shown that large-scale implementation of wind and solar farms can produce significant climate change at continental scales (*10*, *17*). However, in those studies, vegetation is prescribed rather than dynamic—that is, either vegetation types and properties do not respond to the changing climate caused by the large wind and solar farms or the vegetation changes do not feed back onto climate. The lack of vegetation feedbacks could make the modeled climate impacts very different from their actual behavior (*18*, *19*), as vegetation dynamics [e.g., albedo, evapotranspiration, roughness, and leaf area index (LAI)] have been proven to play a key role in the land-climate interaction (*20*). Vegetation feedbacks can either enhance or suppress the initial climate changes triggered by land change (*21*, *22*).

In our study, we used a climate model with dynamic vegetation to investigate the climate impacts of large-scale wind and solar farms installed in the world's largest deserts. We primarily focused on the effect of such large wind and solar farms in the Sahara region (including the most arid parts of the Arabian Desert) and the neighboring Sahel region for several reasons: (i) The Sahara is the largest desert in the world and has a great supply of solar and wind energy. (ii) The Sahara is sparsely inhabited, and thus the development of wind and solar farms would have minimal competition for land surface area against natural and other human land uses, such as agriculture (*15*). (iii) The Sahel is a transition region between desert and wooded savanna and, as such, is highly sensitive to land changes (*18*, *19*, *23*). (iv) Both regions are near Europe and the Middle East, areas with enormous current energy demand, and sub-Saharan Africa, which has a large projected growth in energy demand (see supplementary text). (v) Massive investment in solar and wind generation could promote economic development in the Sahel, one of the poorest regions in the world, as well as provide clean energy for desalination and provision of water for cities and food production (*24*). The wind and solar farms simulated in this study would generate approximately 3 and 79 TW of electrical power, respectively, averaged over a typical year (see supplementary text).

Our results show that the effects of the large-scale wind and solar farms in the Sahara are most significant locally—i.e., at or near the locations of wind and solar farms—with limited remote impacts (Fig. 1). The wind farm causes significant regional warming on near-surface air temperature (+2.16 K), with greater changes in minimum temperature than maximum temperature (+2.36 versus +1.85 K) (fig. S1). This asymmetric temperature impact has been reported in both empirical (*16*) and modeling studies (*14*, *25*). The greater nighttime warming takes place because wind turbines can enhance the vertical mixing and bring down warmer air from above to the lower levels, especially during stable nights (*14*, *26*). Wind farms also increase precipitation as much as +0.25 mm/day, averaged over areas with wind farm installations, which results in the doubling of precipitation compared with the control experiment (0.24 mm/day), particularly in the Sahel region, which features an average increase of +1.12 mm/day (table S1). This is because the increased surface friction reduces wind velocity and the associated Coriolis force, which leads to a more dominant pressure gradient force toward the Saharan heat low that is enhanced by the warming induced by wind farms. This produces surface convergence and upward motion as well as moisture convergence and higher humidity (figs. S3 to S5). The increase in precipitation, in turn, leads to increases in vegetation cover fraction (+0.084), LAI (+0.50 m^2/m^2), and root carbon (+0.08 kgC/m^2) that further reduce surface albedo (figs. S2 and S5). These changes together trigger a positive albedo–precipitation–vegetation feedback (*21*, *22*). Additionally, the recovered vegetation increases evaporation, surface friction, cloud cover (fig. S3), and consequently, precipitation. The increased evaporation, which partially compensates the increased net surface solar radiation, also plays an important role in local precipitation enhancement (*21*). A slight cooling is observed in the wetter Sahel region because recovered vegetation increases evaporation and decreases sensible heat flux. As expected, the increased drag at the surface due to wind turbines reduces wind speed by ~36% (fig. S1).

Impacts of solar farms on temperature and precipitation are markedly similar to those of wind farms in terms of spatial patterns. This is because solar panels directly reduce surface albedo and thus trigger a similar positive albedo–precipitation–vegetation feedback to that of wind farms, and this feedback leads to temperature and precipitation increase. The resulting warming is stronger in maximum temperature

[1]Department of Atmospheric and Oceanic Science, University of Maryland, College Park, MD 20742, USA. [2]Department of Natural Resources and Environmental Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [3]State Key Laboratory of Earth Surface Processes and Resources Ecology, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China. [4]Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742, USA. [5]Department of Physics, University of Maryland, College Park, MD 20742, USA. [6]Earth System Physics Section, Abdus Salam International Centre for Theoretical Physics, Trieste I-34100, Italy. [7]LASG, Institute of Atmospheric Physics, Chinese Academy of Science, Beijing 100029, China.

***Corresponding author. Email: yanli.geo@gmail.com (Y.L.); ekalnay@umd.edu (E.K.); ssm@umd.edu (S.M.)** †Independent researcher.

(+1.28 K) than in minimum temperature (+0.97 K) because albedo reduction mainly affects net radiation during daytime (fig. S1). Compared with the control experiment, a 50% increase in precipitation (+0.13 mm/day) is observed in solar farm locations in the Sahara, and an increase of +0.57 mm/day is recorded in the Sahel (table S1). Unlike the wind farm experiment, the solar farm experiment produces very little change in wind speed (fig. S1).

When wind and solar farms are deployed together in the Sahara, changes in climate are enhanced. The precipitation in the Sahara increases from 0.24 mm/day in the control run to 0.59 mm/day in the case of combined wind and solar farms, a ~150% increase, whereas the temperature increase (+2.65 K) is only slightly larger compared with that for the solar farm alone. Although the absolute amount of precipitation change averaged over the entire Sahara is low from these experiments (up to +0.35 mm/day for the combined wind and solar farms), it should be emphasized that the precipitation impact is not uniform across space. The most substantial precipitation increase occurs in the Sahel, with a magnitude of change between +200 to +500 mm/year (table S1), which is large enough to have major ecological, environmental, and societal impacts.

Our simulations show that both the wind and solar farms in the Sahara contribute to increased precipitation, especially in the Sahel region, through the positive albedo–precipitation–vegetation feedback. This positive feedback is established through different mechanisms for wind and solar farms. For wind farms, the higher surface roughness strengthens low-level convergence, leading to precipitation increase in the Sahara (*27*). For solar farms, the decreased albedo associated with solar panels (i.e., the lower effective albedo of solar panels compared with the sand in the Sahara) results in more absorption of solar radiation and, hence, surface warming, which leads to low pressure at the surface, as well as convergence, rising motion, and consequently, more precipitation (*23*, *28*). The precipitation increase induced by either wind or solar farms, in turn, increases vegetation cover and LAI, leading to further reduction in albedo and increase in roughness, both of which help organize the moisture convergence that drives the change in precipitation. These friction–precipitation–vegetation feedbacks (wind farms) and albedo–precipitation–vegetation feedbacks (wind and solar farms) are known as the Sud (*27*) and Charney mechanisms (*23*, *28*), respectively. To quantify the contribution of these two mechanisms, we carried out additional wind farm experiments in which both mechanisms are present that can separate the climate changes induced by the initial roughness and the subsequent albedo changes due to vegetation feedback (Fig. 2). We found that for the temperature change, roughness and vegetation feedback contribute almost equally (+1.00 versus +1.16 K). The roughness-induced warming occurs because wind reduction weakens the near-surface vertical turbulence transport (*29*). In contrast, for precipitation change, 80% of the increase (+0.20 mm/day) comes from vegetation feedback, whereas roughness plays only a secondary role, except as an initial trigger. These results suggest that the absence of vegetation feedback in the model (*18*, *19*) would considerably underestimate the temperature and precipitation impacts of the large-scale wind farms in the Sahara.

Although wind and solar farms both enhance precipitation in our experiments, solar farms will not necessarily always increase precipitation through albedo changes. The direction of precipitation change is largely determined by the sign of the albedo change before and after a solar farm is installed. More specifically, it depends on the conversion efficiency of the solar panel and the background environment albedo. The precipitation increase in our solar farm experiments is due to the relatively low conversion efficiency of the panels (15%, typical current conversion efficiency for photovoltaic panels), which results in albedo decrease (*30*). However, if solar panel efficiency and the associated effective albedo are high enough to lead to an albedo increase relative to the background environment (as, for example, a 45% efficiency would), the climate impact would be surface cooling with precipitation suppression (fig. S6), similar to the impact of overgrazing in the desert (*23*). Assuming an intermediate conversion efficiency higher than 15% for solar panels (e.g., 30% efficiency) results in

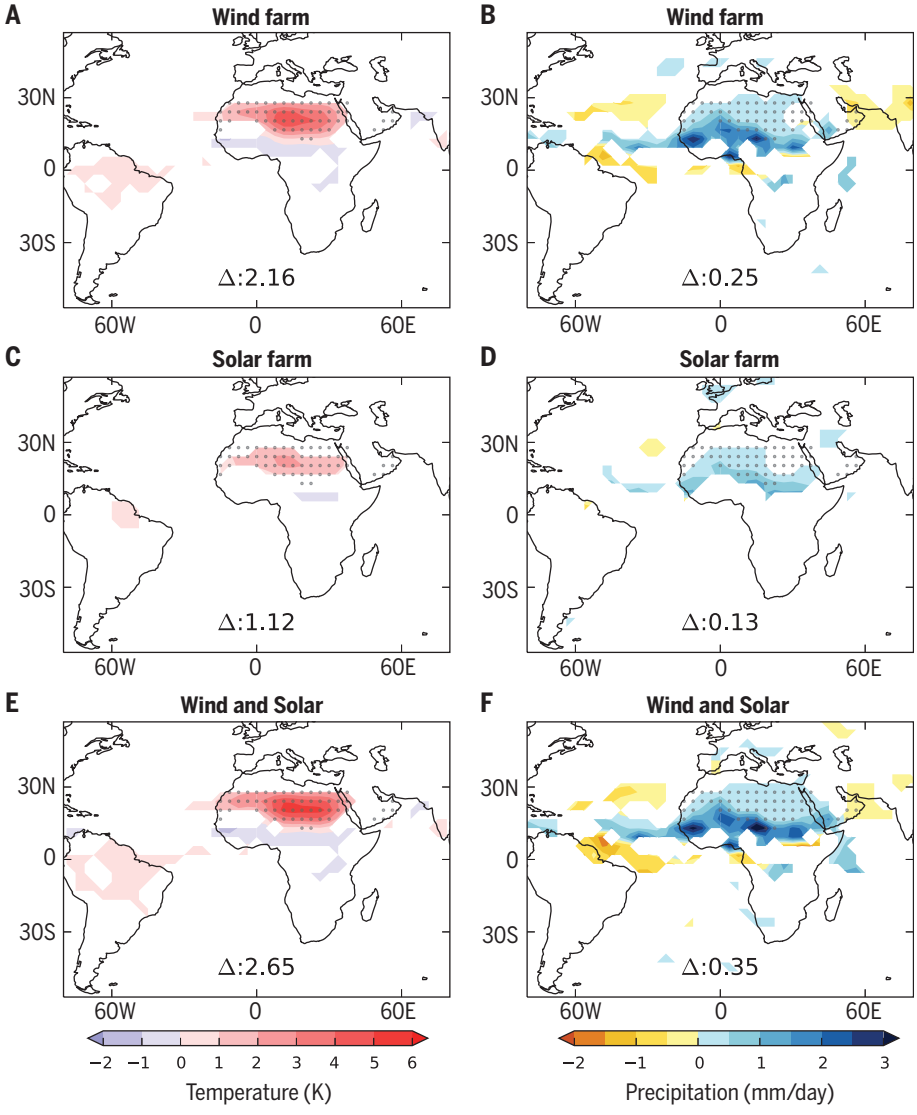


Fig. 1. Impacts of wind and solar farms in the Sahara on mean near-surface air temperature (kelvin) and precipitation (millimeters per day). The impacts of wind farms (**A** and **B**), solar farms (**C** and **D**), and wind and solar farms together (**E** and **F**), respectively, are shown. Only areas where changes are significant at the 95% confidence level (*t* test) are displayed on the map. Gray dots denote the location of wind and/or solar farms. At the bottom of each plot, the number after Δ represents the changes in climate (in either kelvin or millimeters of precipitation per day) averaged over areas covered by wind and solar farms.

negligible albedo change and, thus, insignificant climate impacts (fig. S6).

In this study, we used a model running at a relatively coarse spatial resolution to simulate the impact of wind and solar farms. However, the model has been shown to be capable of capturing the large-scale impacts from changing albedo, roughness, and vegetation responses (*20*–*22*, *31*) and has skills comparable to those of other higher-resolution models in simulating modern climate and multidecadal variability in this region in multiple model intercomparison projects (*32*, *33*). Still, uncertainties remain in the magnitude of climate response and the strength of vegetation feedbacks. The complexity of a global model also limits its ability in capturing the impacts on synoptic and mesoscale weather processes. It is not clear if all of our findings are directly applicable to wind and solar farms with a size much smaller than the model grid. Nevertheless, the temperature impacts of wind and solar farms in our study (i.e., the warming effect of wind farms and the albedo-dependent impact of solar panels) are consistent with those reported in studies conducted at the local scale (*34*, *35*). For the precipitation change, the impact is more uncertain due to its region-specific and scale-dependent nature. We addressed these uncertainties by designing additional experiments (*30*). We found that expanding the wind and solar farms from the Sahara to the world's other deserts does not significantly increase the climate impact (fig. S7). The most significant impact is still concentrated in the Sahara and the surrounding regions, whereas the impact is not significant in many other deserts, due to their scattered geographical distributions, smaller sizes, and weaker changes in albedo (fig. S8). Even in the Sahara, the wind and solar farms impacts also depend on their specific location and spatial distribution, with uneven impacts when deployed with different spatial configurations (i.e., the "checkerboard" and "quarter" wind farm experiments represented in fig. S9). Therefore, to assess the impacts of smaller-scale wind and solar farms installed at specific locations, further studies are required, especially those using more advanced global and regional climate models with higher spatial resolutions (*25*).

Our results obtained from experiments performed with a climate model suggest that, for installations of wind and solar farms with current conversion efficiency in the desert at a scale large enough to power the entire world, the impacts on regional climate would be beneficial rather than detrimental, and the impacts on global mean temperature are still small compared with those induced by CO_2 emission from fossil fuels (*3*, *10*). If carefully planned, these farms could also trigger more precipitation, largely because of a previously overlooked vegetation feedback. This highlights that, in addition to avoiding anthropogenic greenhouse gas emissions from fossil fuels and the resulting warming, wind and solar energy could have other unexpected beneficial climate impacts when deployed at a large scale in the Sahara, where conditions are especially favorable for these impacts. Efforts to build such large-scale wind and solar farms for electricity generation may still face many technological (e.g., transmission, efficiency), socioeconomic (e.g., cost, politics), and environmental challenges, but this goal has become increasingly achievable and cost-effective (*36*) (supplementary text). These results indicate that renewable energy can have multiple benefits for climate and sustainable development and thus could be widely adopted as a primary solution to the challenges of global energy, climate change, and environmental and societal sustainability (*4*).

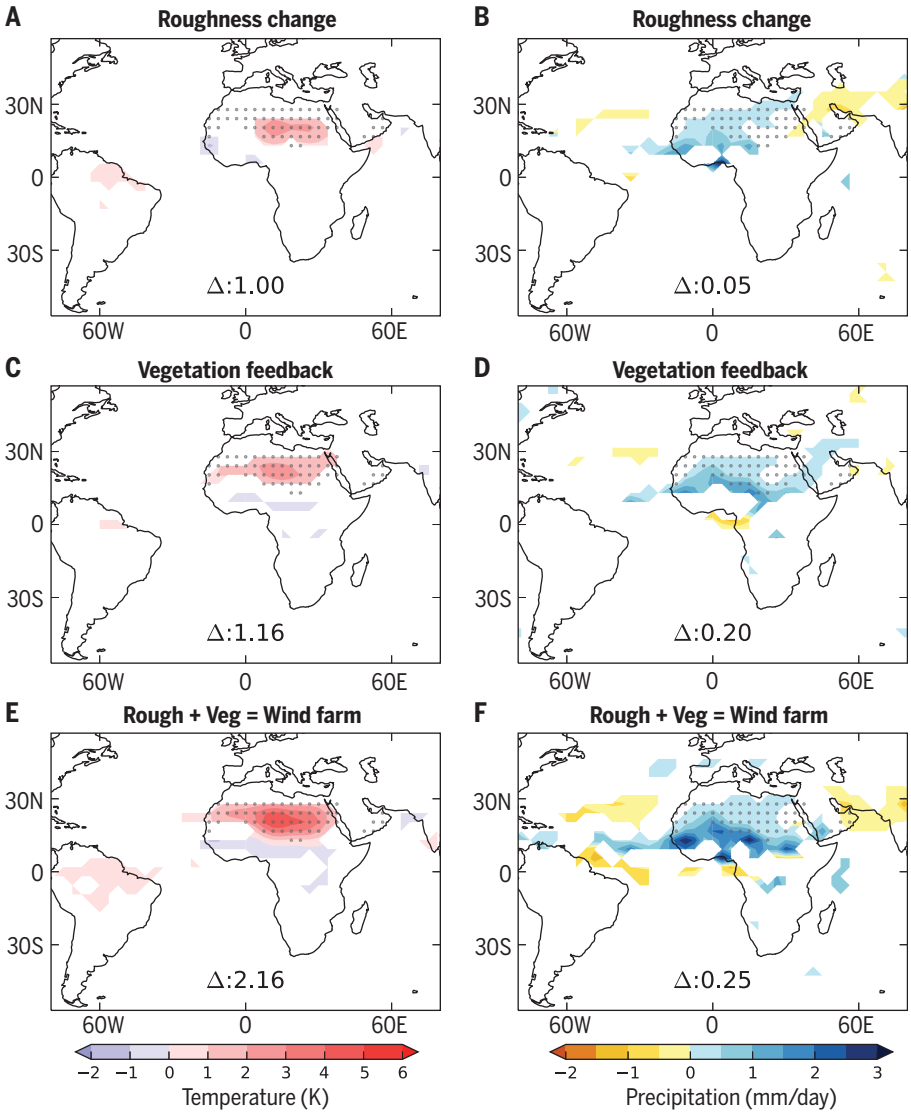


Fig. 2. Relative contributions of roughness change (Rough) and vegetation feedback (Veg) in the climate impacts of wind farms in the Sahara. Contributions in the temperature (**A**, **C**, and **E**) and the precipitation (**B**, **D**, and **F**) impacts are shown. The wind farm impact is produced by the initial roughness of wind turbines and the subsequent albedo changes due to vegetation feedback. At the bottom of each plot, the number after Δ represents the changes in climate (in either kelvin or millimeters of precipitation per day) averaged over areas covered by wind farms.

REFERENCES AND NOTES

1. C. McGlade, P. Ekins, *Nature* **517**, 187–190 (2015).
2. D. T. Shindell, Y. Lee, G. Faluvegi, *Nat. Clim. Change* **6**, 503–507 (2016).
3. S. I. Seneviratne, M. G. Donat, A. J. Pitman, R. Knutti, R. L. Wilby, *Nature* **529**, 477–483 (2016).
4. S. Motesharrei *et al.*, *Natl. Sci. Rev.* **3**, 470–494 (2017).
5. M. Jakob *et al.*, *Nat. Clim. Change* **4**, 961–968 (2014).
6. R. J. Millar *et al.*, *Nat. Geosci.* **10**, 741–747 (2017).
7. A. E. MacDonald *et al.*, *Nat. Clim. Change* **6**, 526–531 (2016).
8. R. J. Barthelmie, S. C. Pryor, *Nat. Clim. Change* **4**, 684–688 (2014).
9. L. M. Miller, F. Gans, A. Kleidon, *Earth Syst. Dyn.* **2**, 1–12 (2011).
10. A. Hu *et al.*, *Nat. Clim. Change* **6**, 290–294 (2015).
11. X. Lu, M. B. McElroy, J. Kiviluoma, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 10933–10938 (2009).
12. K. Marvel, B. Kravitz, K. Caldeira, *Nat. Clim. Change* **3**, 118–121 (2012).

13. M. Z. Jacobson, M. A. Delucchi, M. A. Cameron, B. A. Frew, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 15060–15065 (2015).
14. S. Baidya Roy, S. W. Pacala, R. L. Walko, *J. Geophys. Res. Atmos.* **109**, D19101 (2004).
15. R. R. Hernandez *et al.*, *Renew. Sustain. Energy Rev.* **29**, 766–779 (2014).
16. L. Zhou *et al.*, *Nat. Clim. Change* **2**, 539–543 (2012).
17. D. W. Keith *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 16115–16120 (2004).
18. G. Wang, E. A. B. Eltahir, *Geophys. Res. Lett.* **27**, 795–798 (2000).
19. G. Wang, E. A. B. Eltahir, *Water Resour. Res.* **36**, 1013–1021 (2000).
20. N. Zeng, J. D. Neelin, K. Lau, C. J. Tucker, *Science* **286**, 1537–1540 (1999).
21. F. Kucharski, N. Zeng, E. Kalnay, *Clim. Dyn.* **40**, 1453–1466 (2012).
22. N. Zeng, J. Yoon, *Geophys. Res. Lett.* **36**, L17401 (2009).
23. J. G. Charney, *Q. J. R. Meteorol. Soc.* **101**, 193–202 (1975).
24. D. S. Battisti, R. L. Naylor, *Science* **323**, 240–244 (2009).
25. R. Vautard *et al.*, *Nat. Commun.* **5**, 3196 (2014).
26. S. Baidya Roy, J. J. Traiteur, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 17899–17904 (2010).
27. Y. C. Sud, W. E. Smith, *Boundary-Layer Meteorol.* **33**, 15–49 (1985).
28. J. Charney, W. J. Quirk, S. Chow, J. Kornfield, *J. Atmos. Sci.* **34**, 1366–1385 (1977).
29. C. Wang, R. G. Prinn, *Atmos. Chem. Phys.* **10**, 2053–2061 (2010).
30. Materials and methods are available as supplementary materials.
31. Y. Li *et al.*, *Earth Syst. Dyn.* **7**, 167–181 (2016).
32. A. A. Scaife *et al.*, *Clim. Dyn.* **33**, 603–614 (2009).
33. Y. Xue *et al.*, *Clim. Dyn.* **47**, 3517–3545 (2016).
34. D. Millstein, S. Menon, *Environ. Res. Lett.* **6**, 034001 (2011).
35. H. Taha, *Sol. Energy* **91**, 358–367 (2013).
36. G. C. Wu *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **114**, E3004–E3012 (2017).
37. Y. Li *et al.*, Model simulation data for the climate impacts of large-scale wind and solar farms in the Sahara and the world's deserts, Figshare (2018); doi:10.6084/m9.figshare.6662963.

ACKNOWLEDGMENTS

We thank the University of Maryland and the Univ. of Illinois for supercomputing resources—in particular, the Deepthought2 (http://hpcc.umd.edu) and Bluewaters (www.ncsa.illinois.edu/enabling/bluewaters) supercomputers—made available for conducting the research reported in this paper. We also thank three anonymous reviewers for constructive comments. **Funding:** Y.L. acknowledges support from the National Key R&D Program of China (no. 2017YFA0604701). E.K. and S.M. acknowledge Lev Gandin funding (grant 2956713) provided by G. Brin. **Authors contributions:** E.K., Y.L., and S.M. conceptualized the study; Y.L., E.K., S.M., F.K., and J.R. designed the experiments; Y.L. and E.B. performed the model simulations; F.K., N.Z., and Y.L. developed the UMD-ICTP model version used in this study; Y.L., E.K., S.M., and F.K. analyzed the data, with contributions from other co-authors; and Y.L., S.M., E.K., F.K., and J.R. wrote the manuscript, with discussions and contributions from other co-authors. **Competing interests:** D.K.-D. is a lead research scientist at AWS Truepower, whose work involves forecasting renewable generation for grid operators. **Data and materials availability:** The model simulation data are available at Figshare (*37*).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6406/1019/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S9
Table S1
References (*38*–*55*)

22 November 2017; accepted 1 August 2018
10.1126/science.aar5629

MIGRATION

Is ungulate migration culturally transmitted? Evidence of social learning from translocated animals

Brett R. Jesmer[1,2]*, Jerod A. Merkle[2], Jacob R. Goheen[1], Ellen O. Aikens[1,2], Jeffrey L. Beck[3], Alyson B. Courtemanch[4], Mark A. Hurley[5], Douglas E. McWhirter[4], Hollie M. Miyasaki[5], Kevin L. Monteith[2,6], Matthew. J. Kauffman[7]

Ungulate migrations are assumed to stem from learning and cultural transmission of information regarding seasonal distribution of forage, but this hypothesis has not been tested empirically. We compared the migratory propensities of bighorn sheep and moose translocated into novel habitats with those of historical populations that had persisted for hundreds of years. Whereas individuals from historical populations were largely migratory, translocated individuals initially were not. After multiple decades, however, translocated populations gained knowledge about surfing green waves of forage (tracking plant phenology) and increased their propensity to migrate. Our findings indicate that learning and cultural transmission are the primary mechanisms by which ungulate migrations evolve. Loss of migration will therefore expunge generations of knowledge about the locations of high-quality forage and likely suppress population abundance.

From tropical savannas to the Arctic tundra, the migrations of ungulates (hooved mammals) can span more than 1000 km and are considered among the most awe inspiring of natural phenomena. Migration allows ungulates to maximize energy intake by synchronizing their movements with the emergence of high-quality forage across vast landscapes (*1*). Consequently, migration often bolsters fitness and results in migratory individuals' greatly outnumbering residents (*2*, *3*). Despite their critical importance, migrations are increasingly imperiled by human activities (*4*). Thus, understanding how migrations are developed and maintained is critical for the conservation of this global phenomenon (*5*). Ecologists have long speculated that memory and social learning underlie ungulate migration (*6*–*8*). Bison (*Bison bison*) remember the locations of high-quality forage and transmit such information to conspecifics (*9*), whereas moose (*Alces alces*) and white-tailed deer (*Odocoileus virginianus*) adopt the movement strategies of their mothers (*6*, *7*). Nevertheless, the hypothesis that social learning underlies the development and maintenance of ungulate migration has not been tested with empirical data.

Animal migrations arise through a combination of learned behavior and genetically inherited neurological, morphological, physiological, and behavioral traits (*5*, *10*, *11*). When behavior is primarily a consequence of social learning and persists across generations—a phenomenon known as culture—information is transmitted from generation to generation (*12*). Culture is therefore regarded as a "second inheritance system," analogous to the inheritance of genes that underlie innate behaviors (*13*–*15*). Thus, if social learning is the primary mechanism allowing animals to gain information regarding the seasonal distribution of high-quality forage, cultural transmission may be the principal force by which ungulate migrations have evolved in landscapes conducive to migration.

Ungulate migration is a strategy for exploiting altitudinal, longitudinal, and other topographic gradients of plant phenology that determine forage quality (*16*, *17*). The ability of ungulates to synchronize their movements with phenological waves of nutritious, green plants—a behavior known as "green-wave surfing" (*18*)—can result in migratory movements far beyond an individual's perceptual range (*19*). Ungulates also can surf green waves of forage within year-round ranges, even in the absence of migration (*1*). Green-wave surfing may therefore represent a learned behavior that underlies migration, and such knowledge may accumulate over generations via cultural transmission (*15*, *20*).

Across the American West, many bighorn sheep (*Ovis canadensis*) populations were extirpated in the late 1800s because of market hunting and transmission of disease from domestic sheep (*O. aries*) (Fig. 1). To restore lost populations, wildlife managers translocated individuals from extant, migratory populations into vacant landscapes where extirpated populations once existed (Fig. 1). These individuals therefore had no knowledge about the landscapes into which they were translocated (herein termed "novel landscapes"). Thus, if migration does not stem primarily from a genetically inherited suite of traits, individuals should fail to migrate when first translocated into novel landscapes where migration would be a profitable strategy (*21*).

To test this prediction, we affixed global positioning system (GPS) collars on 129 bighorn sheep sampled from four populations that had been extant for >200 years (herein termed "historical populations") (Fig. 1) and 80 bighorn sheep when the sheep were first translocated into novel landscapes (table S1). We defined migration as movement between distinct seasonal ranges and classified the movement of each collared individual as migratory or resident by using net-squared displacement (*22*) [supplementary materials (SM)]. We then quantified how green waves of forage propagated across individual landscapes (1000 to 3600 km^2) by measuring the date each pixel in a rasterized time series of the normalized difference vegetation index (250-m spatial resolution, 8-day temporal resolution) peaked in forage quality (SM) (*23*). Using this rasterized measure of peak forage quality, we quantified the semivariance (the magnitude of the wave) in the date of peak forage quality across a range of spatial lags (the distance the wave traveled) (SM). Within historical populations, 65 to 100% of individuals migrated, whereas few (<9%; 7 of 80) individuals translocated into novel landscapes migrated (Fig. 2A). The migratory propensity of a population was not related to the magnitude of the green wave or the distance it traveled (fig. S1), meaning that landscape characteristics alone did not explain differences in migratory propensity among populations. The seven translocated individuals that migrated were translocated into existing populations of bighorn sheep (<200 individuals) that had been reestablished three decades before (SM), suggesting cultural transmission of migratory behavior among conspecifics (horizontal transmission). Because individuals from migratory populations failed to migrate when translocated into landscapes where they had no prior experience, genes are unlikely to be the primary agent underlying ungulate migration. Instead, migration may require extended periods of time for social learning and cultural transmission to occur.

To evaluate the hypothesis that green-wave surfing is a learned behavior, we first calculated the surfing ability of each GPS-collared individual as the absolute difference between the day an individual occupied a location and the day forage quality peaked at that location (*23*). We then controlled for the influence that local differences in latitudinal, elevational, and topographical features may have on an individual's ability to surf the green wave (*23*) by comparing observed green-wave surfing ability with those of a "naïve forager" that moved at random and an "omniscient forager" that had complete knowledge of phenological patterns (SM). By doing so, we were able to quantify how much knowledge individuals possessed about local patterns of phenology (fig. S2). We found that the surfing knowledge of bighorn sheep from historical populations was approximately twice that of transplanted

[1]Program in Ecology, Department of Zoology and Physiology, University of Wyoming, Laramie, WY 82071, USA. [2]Wyoming Cooperative Fish and Wildlife Research Unit, Department of Zoology and Physiology, University of Wyoming, Laramie, WY 82071, USA. [3]Department of Ecosystem Science and Management, University of Wyoming, Laramie, WY 82071, USA. [4]Wyoming Game and Fish Department, Jackson, WY 83001, USA. [5]Idaho Department of Fish and Game, Boise, ID 83712, USA. [6]Haub School of Environment and Natural Resources, University of Wyoming, Laramie, WY 82072, USA. [7]U. S. Geological Survey, Wyoming Cooperative Fish and Wildlife Research Unit, Department of Zoology and Physiology, University of Wyoming, Laramie, WY 82071, USA.
*Corresponding author. Email: bjesmer@uwyo.edu

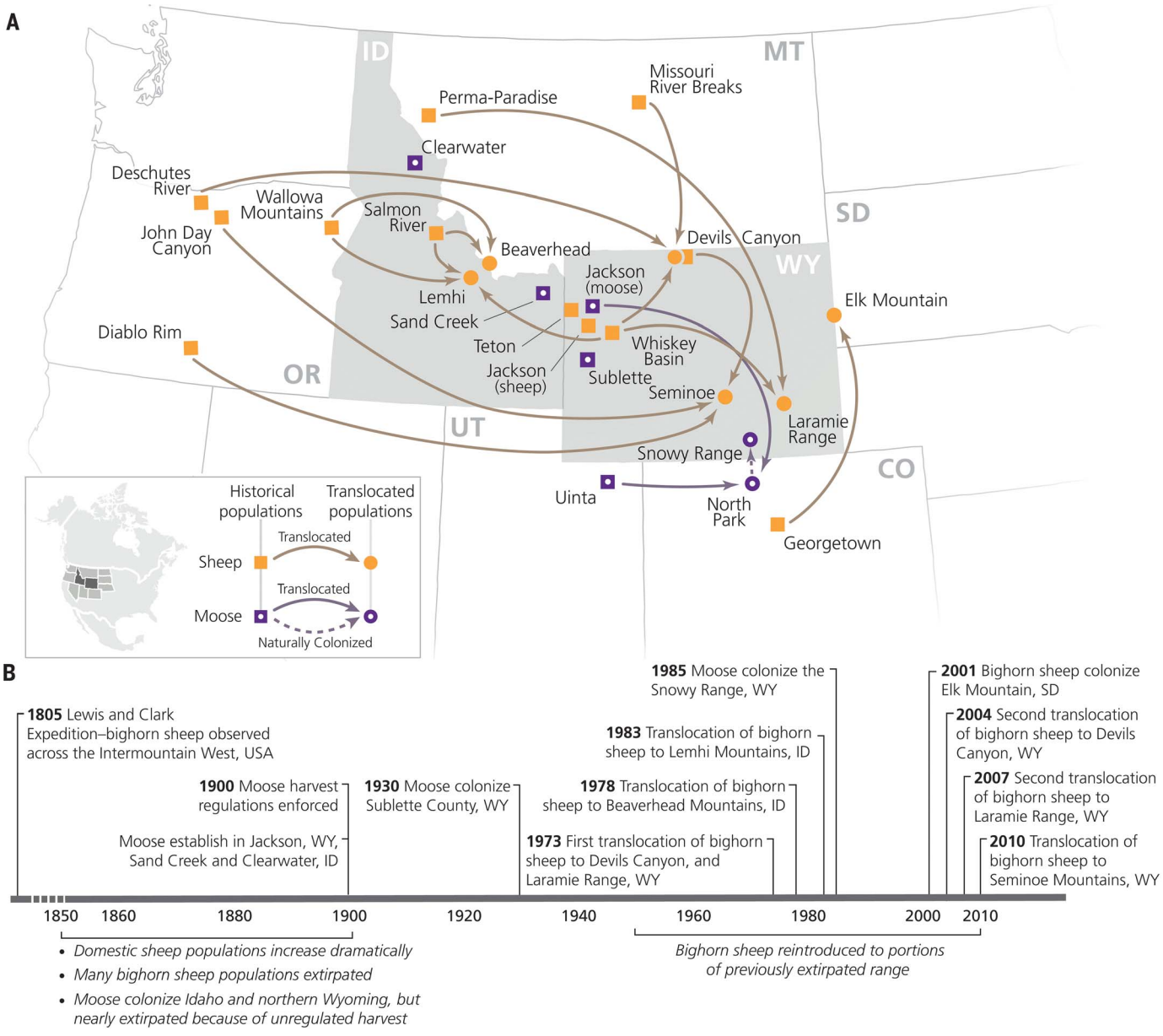


Fig. 1. Bighorn sheep and moose translocation history. (**A**) The subset of historical and translocated populations of bighorn sheep and moose used to assess the cultural basis of ungulate migration. (**B**) Timeline of bighorn sheep and moose translocations as well as other important events in the history of these species since the settlement of western North America by European Americans. See SM for further details about translocation history. (Cartography by InfoGraphics Lab, University of Oregon.)

individuals (Fig. 2B), suggesting that knowledge about local green waves may improve over time as animals learn and culturally transmit information about the seasonal distribution of high-quality forage.

The hypothesis that ungulate migration is established and maintained by cultural transmission predicts that green-wave surfing knowledge and, subsequently, the propensity to migrate should increase as animals learn how to exploit landscapes and transmit that foraging information across generations (vertical transmission of information). To evaluate the influence of vertical transmission on surfing knowledge and migratory propensity, we expanded our analysis to include individuals from four additional populations of bighorn sheep (an additional 58 individuals) and five populations of moose (*Alces alces*; 189 individuals) that were GPS collared ~10 to 110 years after either translocation or natural colonization (Fig. 1, table S1, and SM). We found that the surfing knowledge of both bighorn sheep and moose increased as time since population establishment increased (Fig. 3A). As time passed and bighorn sheep and moose increased their surfing knowledge, their migratory propensities also increased (Fig. 3, B and C). Although population density and migratory propensity are sometimes correlated positively (*24*), migratory propensity did not change with substantial decreases in population density caused by epizootics, habitat loss, and increased predation (*25*, *26*). Together, these results demonstrate that ungulates accumulate knowledge of local phenological patterns over time via the "ratcheting effect," wherein each generation augments culturally transmitted information with information gained from their own experience, a process known as cumulative cultural evolution (*15*, *20*). Cultural transmission therefore acts as a second (nongenetic) inheritance system for ungulates, shaping their foraging and migratory behavior and ultimately providing the primary mechanism by which their migrations have evolved.

Across the globe, anthropogenic barriers have disrupted ungulate migrations, triggered declines

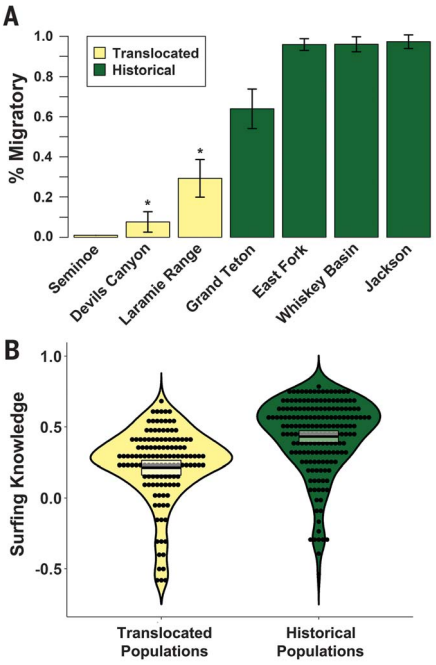


Fig. 2. Migratory propensities and green-wave surfing knowledge of seven translocated and historical populations of bighorn sheep. (**A**) Migratory propensities (±SEM) of bighorn sheep translocated into novel landscapes compared with those of historical populations (>200 years old). Asterisks indicate landscapes where naïve individuals were translocated into populations previously established via translocation ~30 years before. (**B**) Relative to omniscient and naïve foragers on the same landscape, surfing knowledge was lower for translocated (yellow) bighorn sheep than for individuals from historical populations (green). Mean surfing knowledge (black horizontal bars) relative to that of an omniscient forager (set at 1.0) and associated 95% confidence intervals (white boxes) are presented. The surfing knowledge of individuals (black circles) in historical populations was significantly higher than that of translocated individuals (Mann-Whitney *U* Test, $W = 5863$, $P < 0.001$).

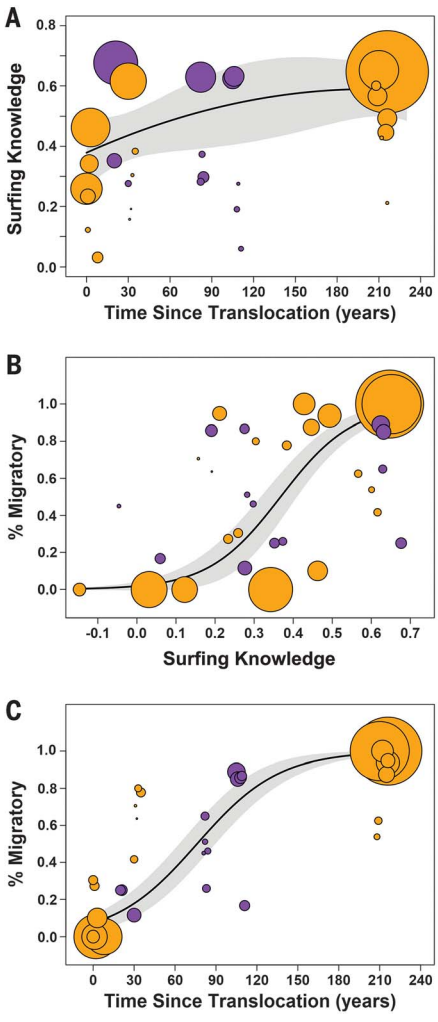


Fig. 3. Green-wave surfing knowledge and migratory propensity over time. (**A**) After translocation, populations of bighorn sheep (orange circles) and moose (purple circles) require decades to learn and culturally transmit information about how to best surf green waves, (**B**) eventually leading to the establishment of migration, which (**C**) takes many generations (the generation time for bighorn sheep and moose is ~7 years). Circles represent estimates of surfing knowledge and migratory propensity for a given population in a given year (i.e., a migratory event). Circle size depicts the amount of confidence (inverse variance) in each estimate. Black lines and gray shaded areas illustrate fitted generalized linear model predictions and their 95% confidence intervals. All relationships are significant at $P < 0.01$.

in population abundance, and even caused local extirpations (*4*). Our results provide empirical evidence that learning and cultural transmission underlie the establishment and maintenance of ungulate migration. Because ungulate migrations stem from decades of social learning about spatial patterns of plant phenology, loss of migration will result in a marked decrease in the knowledge ungulates possess about how to optimally exploit their habitats. Hence, restoring migratory populations after extirpation or the removal of barriers to movement will be hindered by poor foraging efficiency, suppressed fitness, and reduced population performance (*2*, *3*). Thus, conservation of existing migration corridors, stopover sites, and seasonal ranges not only protects the landscapes that ungulates depend on (*27*, *28*); such efforts also maintain the traditional knowledge and culture that migratory animals use to bolster fitness and sustain abundant populations (*13*, *29*).

REFERENCES AND NOTES

1. J. A. Merkle *et al.*, *Proc. R. Soc. London Ser. B* **283**, 20160456 (2016).
2. J. M. Fryxell, J. Greever, A. R. E. Sinclair, *Am. Nat.* **131**, 781–798 (1988).
3. C. M. Rolandsen *et al.*, *Oikos* **126**, 547–555 (2017).
4. G. Harris, S. Thirgood, J. G. C. Hopcraft, J. P. Cromsigt, J. Berger, *Endangered Species Res.* **7**, 55–76 (2009).
5. D. T. Bolger, W. D. Newmark, T. A. Morrison, D. F. Doak, *Ecol. Lett.* **11**, 63–77 (2008).
6. M. E. Nelson, *Can. J. Zool.* **76**, 426–432 (1998).
7. P. Y. Sweanor, F. Sandegren, *J. Appl. Ecol.* **26**, 25–33 (1989).
8. R. B. Boone, S. J. Thirgood, J. G. C. Hopcraft, *Ecology* **87**, 1987–1994 (2006).
9. J. A. Merkle, M. Sigaud, D. Fortin, *Ecol. Lett.* **18**, 799–806 (2015).
10. T. Mueller, R. B. O'Hara, S. J. Converse, R. P. Urbanek, W. F. Fagan, *Science* **341**, 999–1002 (2013).
11. T. Alerstam, *Science* **313**, 791–794 (2006).
12. S. J. Shettleworth, in *Cognition, Evolution, and Behavior* (Oxford University Press, 2010), pp. 417–464.
13. S. A. Keith, J. W. Bull, *Ecography* **40**, 296–304 (2017).
14. A. Whiten, *Nature* **437**, 52–55 (2005).
15. C. Tennie, J. Call, M. Tomasello, *Philos. Trans. R. Soc. London Ser. B* **364**, 2405–2415 (2009).
16. M. Hebblewhite, E. Merrill, G. McDermid, *Ecol. Monogr.* **78**, 141–166 (2008).
17. J. M. Fryxell, *Am. Nat.* **138**, 478–498 (1991).
18. S. A. J. van der Graaf, J. Stahl, A. Klimkowska, J. P. Bakker, R. H. Drent, *Ardea* **94**, 567–577 (2006).
19. C. Bracis, T. Mueller, *Proc. R. Soc. London Ser. B* **284**, 20170449 (2017).
20. T. Sasaki, D. Biro, *Nat. Commun.* **8**, 15049 (2017).
21. K. N. Laland, V. M. Janik, *Trends Ecol. Evol.* **21**, 542–547 (2006).
22. N. Bunnefeld *et al.*, *J. Anim. Ecol.* **80**, 466–476 (2011).
23. E. O. Aikens *et al.*, *Ecol. Lett.* **20**, 741–750 (2017).
24. W. Peters *et al.*, *Ecol. Monogr.* **87**, 297–320 (2017).
25. P. A. Hnilicka *et al.*, in *Northern Wild Sheep and Goat Council 13* (Northern Wild Sheep and Goat Council, 2003), pp. 69–94.
26. B. A. Oates, thesis, University of Wyoming, Laramie, WY (2016).
27. H. Sawyer, M. J. Kauffman, *J. Anim. Ecol.* **80**, 1078–1087 (2011).
28. H. Sawyer *et al.*, *J. Appl. Ecol.* **50**, 68–78 (2013).
29. H. Whitehead, *Learn. Behav.* **38**, 329–336 (2010).
30. B. R. Jesmer *et al.*, Dataset for "Is ungulate migration culturally transmitted?" Dryad (2018); https://doi.org/10.5061/dryad.8165qv5.

ACKNOWLEDGMENTS

Data used in this analysis were collected by, or in collaboration with, biologists at the Wyoming Game and Fish Department and the Idaho Department of Fish and Game who have been working for over half a century to restore and conserve ungulate populations. We thank these biologists for their hard work and dedication. We also thank M. Festa-Bianchet and two anonymous reviewers for providing helpful comments on drafts of the manuscript (see SM for additional acknowledgments). **Funding:** This research was financially supported by the Wyoming Governor's Big Game License Coalition (A.B.C., B.R.J., D.E.M., J.L.B., J.R.G., K.L.M., M.J.K.), the Wyoming Game and Fish Department (D.E.M.), the Idaho Department of Fish and Game (M.A.H., H.M.M.), the Wyoming NASA Space Grant Consortium (B.R.J., J.R.G., M.J.K.), the American Society of Mammalogists (B.R.J.), the Safari Club International Foundation (M.J.K.), the Idaho Safari Club (M.A.H., H.M.M.), the Idaho Transportation Department (M.A.H., H.M.M.), the Bureau of Land Management (M.A.H., H.M.M.), the U.S. Forest Service (M.A.H., H.M.M., A.B.C., J.R.G., M.J.K.), Pittman-Robertson Wildlife Restoration funds (M.A.H., H.M.M.), the Wild Sheep Foundation (H.M.M., M.A.H.), the Wyoming Wild Sheep Foundation (A.B.C., D.E.M., J.L.B., K.L.M., M.J.K.), the Teton Conservation District (A.B.C., M.J.K.), the Grand Teton National Park Foundation (A.B.C., M.J.K.), the Wyoming Wildlife-Livestock Disease Research Partnership (K.L.M.), and the Alces Society (B.R.J.). **Author contributions:** B.R.J., J.A.M., J.R.G., and M.J.K. conceived the study and wrote and revised the manuscript; B.R.J., E.O.A., and J.A.M. analyzed the data; and all coauthors assisted with data collection. **Competing interests:** All authors declare that they have no competing interests. **Data and materials availability:** Data reported in this paper are archived in Dryad (*30*).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6406/1023/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 and S2
Tables S1 to S3
References (*31*–*78*)

26 January 2018; accepted 6 August 2018
10.1126/science.aat0985

PLANT SCIENCE

A single transcription factor promotes both yield and immunity in rice

Jing Wang[1]*†, Lian Zhou[1]*, Hui Shi[1]*, Mawsheng Chern[2]*, Hong Yu[3]*, Hong Yi[1], Min He[1], Junjie Yin[1], Xiaobo Zhu[1], Yan Li[1], Weitao Li[1], Jiali Liu[1], Jichun Wang[1], Xiaoqiong Chen[1], Hai Qing[1], Yuping Wang[1], Guifu Liu[3], Wenming Wang[1], Ping Li[1], Xianjun Wu[1], Lihuang Zhu[3], Jian-Min Zhou[3,4], Pamela C. Ronald[2], Shigui Li[1], Jiayang Li[3,4]†, Xuewei Chen[1]†

Plant immunity often penalizes growth and yield. The transcription factor Ideal Plant Architecture 1 (IPA1) reduces unproductive tillers and increases grains per panicle, which results in improved rice yield. Here we report that higher IPA1 levels enhance immunity. Mechanistically, phosphorylation of IPA1 at amino acid Ser^{163} within its DNA binding domain occurs in response to infection by the fungus *Magnaporthe oryzae* and alters the DNA binding specificity of IPA1. Phosphorylated IPA1 binds to the promoter of the pathogen defense gene *WRKY45* and activates its expression, leading to enhanced disease resistance. IPA1 returns to a nonphosphorylated state within 48 hours after infection, resuming support of the growth needed for high yield. Thus, IPA1 promotes both yield and disease resistance by sustaining a balance between growth and immunity.

Plant growth is usually slowed by an active immune response, resulting in yield penalties for crops fighting pathogens (*1*, *2*). Plants without an active immune response may grow faster but will easily succumb to various diseases (*3*). Various proteins control the growth-immunity trade-off. For example, *Arabidopsis* BRASSINAZOLE-RESISTANT 1 (BZR1) and HOMOLOG OF BRASSINOSTEROID ENHANCED EXPRESSION 2 (BEE2) INTERACTING WITH IBH1 (HBI1) promote plant growth but suppress immunity (*4*, *5*). Conversely, transcription factors TL1-BINDING FACTOR 1 (TBF1) and WRKY45 enhance immunity but inhibit plant growth (*6*, *7*). Breeding practice has selected crop varieties with high yield and disease resistance. A better balance between growth and immunity is supported by various genes, including a natural allele of the *Broad-Spectrum Resistance-Digu 1* (*Bsr-d1*) transcription factor (*8*); a chemically induced allele of *Broad-Spectrum Resistance-Kitaake 1* (*Bsr-k1*), which encodes an RNA binding protein (*9*); a nucleotide-binding oligomerization domain–like receptor (NLR) pair, *Pyricularia-Gumei Resistant* and *Pyricularia-Gumei Susceptible* (*PigmR* and *PigmS*) (*10*); and an artificial, pathogen-inducible cassette containing *Nonexpressor of Pathogenesis-Related genes 1* (*NPR1*) or *snc1* (*suppressor of npr1-1, constitutive 1*) (*11*). However, there are no reports showing that a single protein can positively promote yield and disease resistance.

Rice feeds half of the world's population; improved yields would help sustain the food supply needed for the growing world population. Grain yield depends on the number of productive tillers per plant, the number of grains per panicle, and grain weight (*12*). The *Ideal Plant Architecture 1* (*IPA1*) gene encodes a SQUAMOSA promoter binding protein–like (SPL) transcription factor, also known as OsSPL14, which activates yield-related genes, including *Dense and Erect Panicle 1* (*DEP1*), leading to plants with fewer unproductive tillers and more grains per panicle, supporting higher yield (*13*–*15*). The *ipa1-1D* allele carries a mutation at the miR156 and miR529 target sites, releasing suppression by miR156 and miR529 and leading to higher *IPA1* RNA and protein levels (*13*, *16*). Though *ipa1-1D* plants have been demonstrated to have 10% higher yields in extensive field trials (*13*), it was not known if the improved yield would persist when plants faced pathogen challenges. We directly tested the yield of *ipa1-1D* plants under challenge with *Magnaporthe oryzae*, which causes the devastating rice blast disease. We conducted field tests in three consecutive years by using isogenic rice lines developed in two rice varieties (R320 and R441) and found that *ipa1-1D* plants R320$^{ipa1\text{-}1D}$ and R441$^{ipa1\text{-}1D}$ had yields 10.1 to 13.3% higher under normal field conditions without blast disease and 30.7 to 48.2% higher under high blast disease pressure than controls R320 and R441, respectively (Fig. 1, A to D). The 10 to 13% yield increase under normal conditions is consistent with previous reports (*13*). *ipa1-1D*–mediated yield increase is greater under blast disease pressure, indicating that *IPA1* may also improve resistance to *M. oryzae*. To test this hypothesis directly, we generated *IPA1* overexpression [IPA1–green fluorescent protein (IPA1-GFP)] plants and plants with *IPA1* expression reduced by RNA interference (RNAi) (*13*). *IPA1* overexpression lines showed enhanced resistance and *IPA1* RNAi lines showed higher susceptibility to multiple isolates of *M. oryzae* in both detached leaves and spray-inoculated plants (Fig. 1, E to G, and figs. S1 to S3).

Because *IPA1* RNA and protein levels do not change upon *M. oryzae* infection (Fig. 2A), we investigated whether IPA1 protein becomes phosphorylated in *ipa1-1D* plants upon *M. oryzae* infection. The phosphorylated IPA1 protein was separated from nonphosphorylated IPA1 in a gel containing Phos-tag and detected with an IPA1 polyclonal antibody (Ab). Phosphorylated IPA1 protein starts to accumulate at 3 hours post-infection (hpi), peaks at 6 to 12 hpi, and then subsides to near normal levels within 48 hpi (Fig. 2B and fig. S4). A conserved serine residue exists among different SPL proteins and has been suggested as a phosphorylation site necessary for the transcriptional activity of SPL proteins (*17*). We therefore generated a polyclonal Ab against a 14–amino acid peptide containing phosphorylated Ser^{163} (S163-P) (fig. S5A). The Ab (αIPA1S163-P) recognizes IPA1 containing S163-P [IPA1(S163-P)] with high specificity. Changing S163 to alanine (S163A), which removes the ability of IPA1 to be phosphorylated, abolished IPA1 recognition by αIPA1S163-P (fig. S5B). In samples from *M. oryzae*–infected *ipa1-1D* plants, αIPA1S163-P detected an IPA1 phosphorylation pattern similar to that detected by using Phos-tag, peaking at 12 hpi with ~3-fold enrichment of the IPA1 (S163-P) protein (Fig. 2C). Wild-type plants displayed a similar but weaker phosphorylation response (~2-fold enrichment) upon *M. oryzae* infection (fig. S6). These results indicate that IPA1 S163 becomes phosphorylated upon *M. oryzae* infection in a manner similar to the overall phosphorylation pattern of IPA1.

We next used chromatin immunoprecipitation sequencing (ChIP-seq) to identify genes up-regulated by IPA1 in IPA1-GFP plants (*15*). We found that defense-related genes, including transcription factor *WRKY45*, were up-regulated (fig. S7A). *WRKY45* is required for benzothiadiazole-inducible and NLR protein–mediated immunity to *M. oryzae*, and its elevated expression enhances resistance in rice (*18*, *19*). Thus, elevated *WRKY45* expression may mediate the enhanced pathogen resistance in *IPA1* overexpression plants. Two SPL binding sites containing GTAC sequences were identified in the *WRKY45* promoter (fig. S7B). As expected, probes carrying each of these two sites bound to IPA1 in vitro and in vivo (fig. S7, C and D). Overexpression of *IPA1* in IPA1-GFP plants increased *WRKY45* expression (fig. S7E), indicating that IPA1 activates the *WRKY45* promoter. Moreover, *WRKY45* induction by *M. oryzae* infection was enhanced in *ipa1-1D* plants but reduced

[1]State Key Laboratory of Hybrid Rice, Key Laboratory of Major Crop Diseases and Collaborative Innovation Center for Hybrid Rice in Yangtze River Basin, Rice Research Institute, Sichuan Agricultural University at Wenjiang, Chengdu, Sichuan 611130, China. [2]Department of Plant Pathology and the Genome Center, University of California, Davis, CA 95616, USA. [3]State Key Laboratory of Plant Genomics and National Center for Plant Gene Research (Beijing), Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China. [4]University of the Chinese Academy of Sciences, Beijing 100049, China.
*These authors contributed equally to this work.
†Corresponding author. Email: xwchen88@163.com (X.C.); jyli@genetics.ac.cn (J.L.); jingwang406@sicau.edu.cn (J.W.)

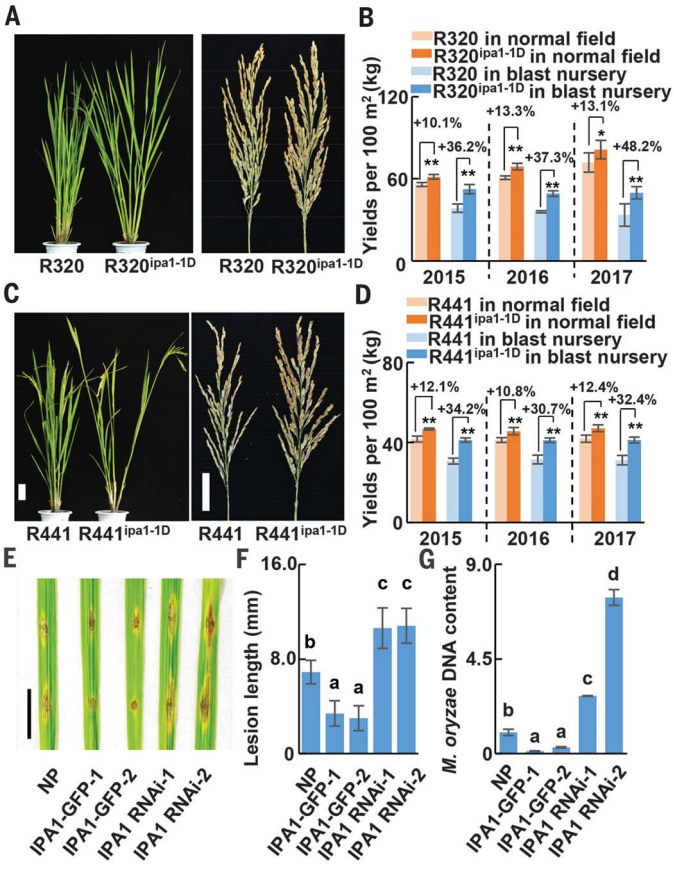


Fig. 1. Elevated IPA1 levels enhance resistance to *M. oryzae*. (**A**) The *ipa1-1D* allele alters plant architecture. *ipa1-1D* was introduced into Shuhui527. The *ipa1-1D* line R320$^{ipa1\text{-}1D}$ and the R320 control were selected from BC_2F_8 progeny. Whole plants and panicles are displayed. Scale bars, 5 cm. (**B**) Yields of R320$^{ipa1\text{-}1D}$ and R320 were tested in the presence of high blast disease pressure (blast nursery) or the absence of blast disease (normal field). Field tests were conducted in 2015, 2016, and 2017. Each data set contained three plots. *P < 0.05; **P < 0.01. (**C**) Same as for (A), except that *ipa1-1D* was introduced into Chenghui3203 and R441$^{ipa1\text{-}1D}$ and R441 were selected. (**D**) Same as for (B), except that R441$^{ipa1\text{-}1D}$ was tested against R441. In (B) and (D), the percent difference was calculated by comparing with the corresponding control. (**E** to **G**) *IPA1* overexpression (IPA1-GFP) enhances resistance and RNAi reduces resistance to *M. oryzae*. Wild-type Nipponbare (NP), IPA1-GFP, and RNAi plants were inoculated with *M. oryzae* isolate Zhong10-8-14. (E) Photographs of lesions. Scale bar, 1 cm. (F) Lesion lengths (n = 10 lesions). (G) *M. oryzae* population (n = 3 repeats). Values are means ± SD. Different letters indicate significant differences determined by the Tukey-Kramer test.

in *IPA1* RNAi plants compared with wild-type plants (Fig. 3A and fig. S8). These results demonstrate the involvement of *WRKY45* in IPA1-mediated immunity to *M. oryzae*. In contrast, expression of *DEP1*, a yield-related gene promoted by IPA1, was suppressed by *M. oryzae* infection (Fig. 3B), suggesting that *M. oryzae*–triggered phosphorylation may change IPA1 DNA binding activity.

To test this hypothesis, we first created two IPA1 mutants, IPA1(S163A) and IPA1(S163D). S163A abolishes phosphorylation capacity, and Ser^{163}→Asp (S163D) mimics S163-P without altering IPA1 nuclear localization (fig. S9). IPA1 (S163D) had reduced binding to the GTAC sites on both *DEP1* and *WRKY45* promoters in electrophoretic mobility shift assays (EMSAs); IPA1 (S163A) had little effect (Fig. 3C and fig. S10A). In addition to GTAC, we previously identified another motif, TGGGCC/T, enriched in the IPA1 ChIP-seq assay (*15*). IPA1(S163D) increased binding to the cis motif TGGGCC, present in the *WRKY45* promoter but not in the *DEP1* promoter (Fig. 3C and fig. S10B). IPA1(S163A) remained similar to the wild type (Fig. 3C). These results were confirmed by ChIP–quantitative polymerase chain reaction (PCR) assays, where IPA1(S163D) pulled down more *WRKY45* TGGGCC sequence (Fig. 3D). Thus, IPA1(S163D) preferentially binds to the TGGGCC site in the *WRKY45* promoter but does not bind the *DEP1* promoter.

To confirm the effects of differential DNA binding ability in planta upon S163 phosphorylation, we overexpressed IPA1(S163D) (labeled S163D-OE) and IPA1(S163A) (labeled S163A-OE) in rice and assessed their effects on immunity. IPA1 S163D-OE plants had smaller lesions and *M. oryzae* populations, whereas S163A-OE had no significant effects (Fig. 4, A to D, and fig. S11). Consistent with this result, *WRKY45* RNA levels were elevated six- to sevenfold in S163D-OE plants; in

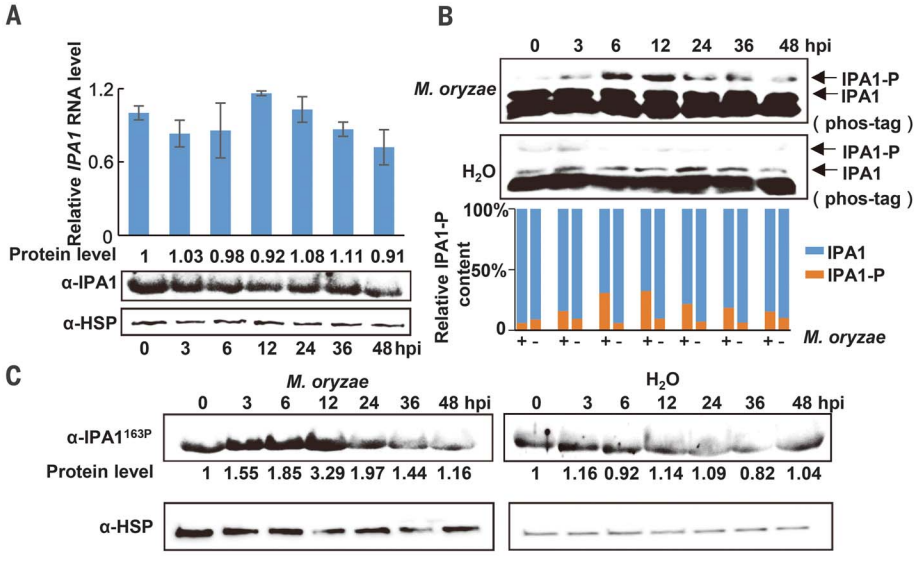


Fig. 2. *M. oryzae* infection induces phosphorylation of IPA1 at S163. (**A**) IPA1 RNA and protein levels are not significantly affected by *M. oryzae* infection. The IPA1 RNA (top) and protein (bottom) levels were assessed at different hpi with *M. oryzae*. The IPA1 protein level was quantitated and normalized to the heat shock protein (HSP) level; the value at time zero was set as one. Error bars indicate SD. (**B**) Phosphorylation of IPA1 is induced upon *M. oryzae* infection. Leaves were collected at different hpi with *M. oryzae* (top) or after treatment with H_2O as a control (middle). Phosphorylated and nonphosphorylated IPA1 proteins were separated on a Phos-tag gel, detected by IPA1 Ab, and quantitated by densitometry, and percentages were calculated (bottom). (**C**) *M. oryzae* infection enhances IPA1 phosphorylation at S163. Immunoblots were probed with an Ab specifically recognizing IPA1 phosphorylated at S163 (IPA1^{163P}) after *M. oryzae* (left) or H_2O (right) treatment. IPA1^{163P} protein amounts were quantitated by densitometry and normalized to the HSP level. The value at time zero was set as one.

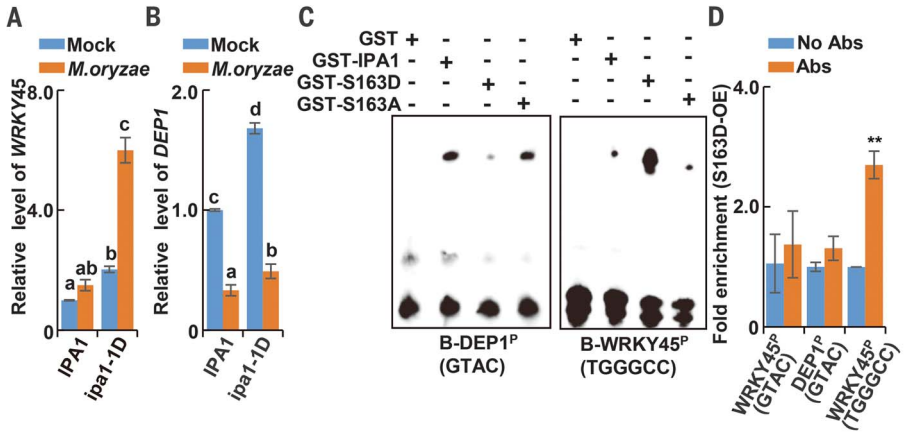


Fig. 3. IPA1(S163D) preferentially binds to the TGGGCC site in the *WRKY45* promoter. (**A**) *M. oryzae* infection induces higher *WRKY45* expression in *ipa1-1D* plants than in wild-type plants. (**B**) *M. oryzae* infection represses *DEP1* expression. In (A) and (B), RNA levels were determined by real-time PCR. (**C**) IPA1(S163D), a mimic of IPA1 phosphorylation at S163, changes DNA binding specificity. IPA1(S163D) reduces binding to the GTAC site in the *DEP1* promoter (left) and enhances binding to the TGGGCC site in the *WRKY45* promoter (right) in EMSAs. GST, glutathione *S*-transferase; B-DEP1^{P} and B-WRKY45^{P}, biotin-labeled *DEP1* and *WRKY45* promoters. (**D**) IPA1(S163D) preferentially binds the *WRKY45* TGGGCC site in a ChIP assay. Values are means ± SD (n = 3 repeats) in (A), (B), and (D). Letters indicate significant differences determined by the Tukey-Kramer test. $^{**}P < 0.01$.

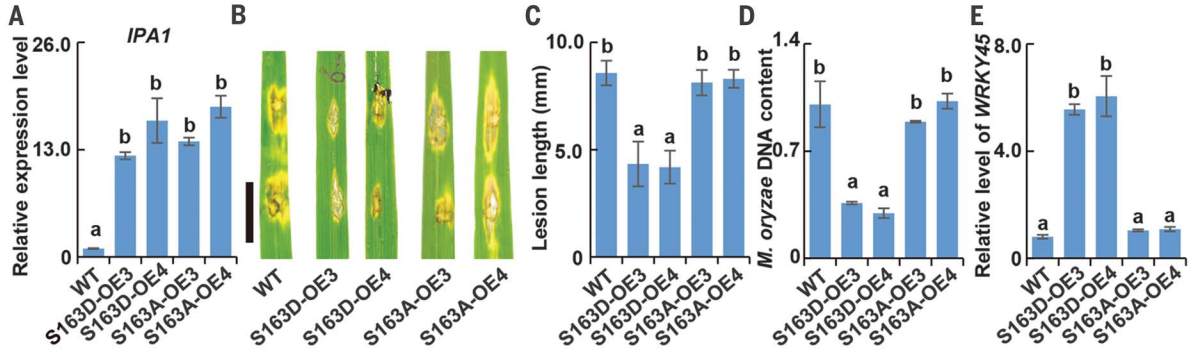


Fig. 4. IPA1(S163D) but not IPA1(S163A) overexpression induces *WRKY45* expression and enhances resistance to *M. oryzae*. IPA1(S163D) (labeled S163D-OE3 and S163D-OE4), IPA1(S163A) (labeled S163A-OE3 and S163A-OE4), and wild-type (WT) plants were inoculated with *M. oryzae*. (**A**) *IPA1* RNA levels. (**B**) Lesion pictures. Scale bar, 1 cm. (**C**) Lesion lengths. (**D**) *M. oryzae* population postinfection. (**E**) *WRKY45* RNA levels. Values are means ± SD. n = 3 repeats in (A), (D), and (E); n = 10 lesions in (C). Different letters indicate significant differences determined by the Tukey-Kramer test.

contrast, S163A-OE activated *DEP1* expression (fig. S12) but not *WRKY45* expression (Fig. 4E). These results demonstrate that IPA1(S163D) but not IPA1(S163A) induces *WRKY45* expression and activates immunity, suggesting that the phosphorylation of IPA1 S163 is critical for the ability to induce *WRKY45* expression and enhance immunity. Moreover, *WRKY45* up-regulation and IPA1 S163 phosphorylation induced by *M. oryzae* infection in *ipa1-1D* plants followed the same pattern (Fig. 2 and fig. S13), further supporting the importance of S163 phosphorylation.

In summary, we discovered that a single protein, IPA1, promotes both yield and disease resistance, and we uncovered its mechanism for controlling two different biological processes. Here, we propose a model for IPA1 function in *ipa1-1D* plants (fig. S14). In the absence of a pathogen, IPA1 is nonphosphorylated at S163 and binds to and activates the *DEP1* promoter, promoting plant growth and yield. Upon pathogen attack, IPA1 becomes phosphorylated at S163. Phosphorylated IPA1 changes DNA binding specificity, switching to bind to the TGGGCC site in the *WRKY45* promoter, and activates *WRKY45* expression, leading to enhanced immunity to *M. oryzae*. Because constitutive phosphorylation of IPA1 would reduce yield (fig. S12), IPA1 returns to the nonphosphorylated state that activates the genes needed for growth and high yield within 48 hpi. In this way, inducible phosphorylation of IPA1 promotes plant growth in the absence of a pathogen and promotes immunity upon pathogen attack. Wild-type plants follow the same phosphorylation pattern for IPA1 as *ipa1-1D* plants, albeit at a lower magnitude (fig. S6). Because *ipa1-1D* plants carry higher levels of nonphosphorylated IPA1 for yield and higher levels of phosphorylated IPA1 protein upon pathogen attack, *ipa1-1D* plants have both improved grain yield and improved immunity. Furthermore, changing DNA binding specificity via phosphorylation of an amino acid to nimbly control different outcomes may prove to be a widespread phenomenon.

REFERENCES AND NOTES

1. J. K. Brown, *Trends Genet.* **19**, 667–671 (2003).
2. R. Nelson, T. Wiesner-Hanks, R. Wisser, P. Balint-Kurti, *Nat. Rev. Genet.* **19**, 21–33 (2018).
3. H. Shi *et al.*, *Plant Cell* **25**, 1143–1157 (2013).
4. M. Fan *et al.*, *Plant Cell* **26**, 828–841 (2014).
5. R. Lozano-Durán *et al.*, *eLife* **2**, e00983 (2013).
6. K. M. Pajerowska-Mukhtar *et al.*, *Curr. Biol.* **22**, 103–112 (2012).
7. S. Goto *et al.*, *Plant Biotechnol. J.* **13**, 753–765 (2015).
8. W. Li *et al.*, *Cell* **170**, 114–126.e15 (2017).
9. X. Zhou *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 3174–3179 (2018).
10. Y. Deng *et al.*, *Science* **355**, 962–965 (2017).
11. G. Xu *et al.*, *Nature* **545**, 491–494 (2017).
12. Y. Wang, J. Li, *Plant Mol. Biol.* **59**, 75–84 (2005).
13. Y. Jiao *et al.*, *Nat. Genet.* **42**, 541–544 (2010).
14. K. Miura *et al.*, *Nat. Genet.* **42**, 545–549 (2010).
15. Z. Lu *et al.*, *Plant Cell* **25**, 3743–3759 (2013).
16. L. Wang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 15504–15509 (2015).
17. R. P. Birkenbihl, G. Jach, H. Saedler, P. Huijser, *J. Mol. Biol.* **352**, 585–596 (2005).
18. M. Shimono *et al.*, *Plant Cell* **19**, 2064–2076 (2007).
19. H. Inoue *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 9577–9582 (2013).

ACKNOWLEDGMENTS

We thank Y. Qian (Biogle) for rice transformation. **Funding:** This work was supported by NSFC (31171622, 31571994, 31788103, 31401351, 31601290, and 31701779), NKRDPC (2016YFD0100600), PNCETC (NECT-13-0920), the NSF (1237975), the NIH (GM59962), and NIFA (2017-67013-26590). **Author contributions:** Jin.W., J.Li, and Xu.C. conceived and designed the experiments. Jin.W., L.Zho., H.S., H.Yu, H.Yi, M.H., X.Z., Y.L., W.L., J.Liu, and Xi.C. performed experiments with phenotypic and biochemical assays. J.Y., Jic.W., Y.W., G.L., and H.Q. contributed to rice materials. Jin.W., M.C., H.Yu, W.W., P.L., X.W., and S.L. collected data. Jin.W., M.C., L.Zhu, J.-M.Z., P.C.R., J.Li, and Xu.C. analyzed the data and wrote the manuscript. **Competing interests:** None declared. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper or the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6406/1026/suppl/DC1
Materials and Methods
Figs. S1 to S14
Table S1
References (*20*–*27*)

12 June 2018; accepted 16 July 2018
10.1126/science.aat7675

FERTILIZATION

The Ly6/uPAR protein Bouncer is necessary and sufficient for species-specific fertilization

Sarah Herberg[1], Krista R. Gert[1], Alexander Schleiffer[1,2], Andrea Pauli[1]*

Fertilization is fundamental for sexual reproduction, yet its molecular mechanisms are poorly understood. We found that an oocyte-expressed Ly6/uPAR protein, which we call Bouncer, is a crucial fertilization factor in zebrafish. Membrane-bound Bouncer mediates sperm-egg binding and is thus essential for sperm entry into the egg. Remarkably, Bouncer not only is required for sperm-egg interaction but is also sufficient to allow cross-species fertilization between zebrafish and medaka, two fish species that diverged more than 200 million years ago. Our study thus identifies Bouncer as a key determinant of species-specific fertilization in fish. Bouncer's closest homolog in tetrapods, SPACA4, is restricted to the male germline in internally fertilizing vertebrates, which suggests that our findings in fish have relevance to human biology.

Fertilization, whereby two gametes fuse to form the single-cell zygote in sexually reproducing organisms, is highly efficient yet species-restricted. This strategy ensures reproductive success and the survival of distinct species. However, the means by which nature has fulfilled these seemingly contradictory requirements, particularly at the molecular level, have remained a mystery. The only vertebrate proteins known so far to be essential for sperm-egg binding are the sperm-expressed IZUMO1 (*1*, *2*) and the egg membrane proteins JUNO (*3*) and CD9 (*4*–*6*). Binding of IZUMO1 to JUNO mediates adhesion between sperm and egg in mammals (*1*–*3*, *7*), whereas the role of CD9 in this process remains unclear. Although in vitro binding assays show that human IZUMO1 binds more efficiently to human JUNO than to mouse JUNO (*8*), an in vivo function in mediating species specificity has not been identified for any of these factors.

To identify factors required for fertilization in vertebrates, we examined our collection of predicted protein-coding genes (*9*) that are expressed in zebrafish oocytes and/or testis. A single-exon gene stood out because of its high expression in zebrafish oocytes (Fig. 1A) and the presence of homologous sequences in other vertebrates. On the basis of its loss-of-function phenotype (see below), we named this gene *bouncer* (*bncr*) in reference to the colloquial name of a security guard at a bar. Although *bouncer* lacks any gene annotation in the newest zebrafish genome release (GRCz11), our RNA sequencing (RNA-seq) and in situ hybridization analyses (fig. S1, A and B), ribosome profiling data (*9*–*11*), and cap analysis gene expression (CAGE)–based transcription start site analysis (*12*) suggested that *bouncer* is a maternal transcript that generates a mature 80–amino acid glycosylphosphatidylinositol (GPI)–anchored protein (Fig. 1A). Consistent with two predicted N-glycosylation sites (Fig. 1A), a Bouncer-specific antibody detected glycosylated Bouncer in the egg (Fig. 1B).

A protein domain search classified Bouncer as a member of the Ly6/uPAR (Ly6/urokinase-type plasminogen activator receptor) protein superfamily, which includes proteins as diverse as toxins, immunoregulators, and cell surface receptors (*13*). This protein family is characterized by a 60– to 80–amino acid domain containing 8 to 10 highly conserved cysteines that form a three-finger structure (Fig. 1, A and C, and fig. S2A). Apart from the cysteines, other amino acids have diverged substantially within this protein superfamily (Fig. 1C and fig. S2, A and B). BLASTP searches with zebrafish Bouncer and phylogenetic sequence analyses suggested that SPACA4 is the closest homolog in mammals, reptiles, and amphibians (Fig. 1C, fig. S2, A to C, data S1 and S2, and table S1). Human SPACA4/SAMP14 (sperm acrosome membrane-associated protein 4/sperm acrosomal membrane protein 14) was originally identified in a proteomics study as a sperm acrosomal protein, and in vitro experiments implied a possible function in fertilization (*14*). However, the in vivo function and importance of SPACA4 are unknown.

Intrigued by our finding that zebrafish Bouncer is expressed in oocytes and that its closest homolog in humans was reported to be expressed in sperm (*14*), we analyzed the expression patterns of other Bouncer/SPACA4 homologs. We found that externally fertilizing vertebrates (e.g., fish and amphibians) show oocyte-restricted expression, whereas all internally fertilizing vertebrates analyzed (e.g., reptiles and mammals) show testis-specific expression (Fig. 1C and fig. S2A). Together, these results identified Bouncer and SPACA4 as homologs with opposing sex-specific, germline-restricted expression patterns in externally versus internally fertilizing vertebrates.

To investigate the function of Bouncer, we used CRISPR/Cas9 to generate *bouncer* knockout zebrafish. We established a stable mutant line with a *bouncer* allele carrying a 13-nucleotide deletion, which abolishes the production of mature Bouncer protein (Fig. 1B and fig. S3A). Incrosses of *bouncer* heterozygous ($bncr^{+/-}$) fish gave rise to homozygous mutant adults ($bncr^{-/-}$) at a Mendelian ratio of ~25%, which suggests that Bouncer is not essential for development. However, in vivo mating experiments showed that only 7 of 3024 eggs (0.11%) derived from $bncr^{-/-}$ females developed into cleavage-stage embryos, as opposed to the majority of eggs from wild-type or $bncr^{+/-}$ females or wild-type eggs fertilized by $bncr^{-/-}$ males (Fig. 1, D and E, and fig. S3, A and B). Notably, female near-sterility was fully rescued by ubiquitous expression of transgenic untagged or green fluorescent protein (GFP)–tagged Bouncer (Fig. 1, D and E, and fig. S3, A and C), which confirms that the observed defect was indeed due to the lack of Bouncer protein. Ubiquitous expression of a Bouncer mutant that cannot be glycosylated (GFP-$Bncr^{N32A,N84A}$; fig. S3, A and D) also fully rescued female near-sterility (Fig. 1E); this finding demonstrates that glycosylation of Bouncer does not contribute to its function. Thus, oocyte-expressed Bouncer protein is necessary for efficient reproduction in zebrafish.

Zebrafish eggs are activated upon contact with spawning medium, independently of the presence of sperm. Egg activation appeared unaffected in eggs from $bncr^{-/-}$ females, as was evident by normal elevation of the chorion (the outer protective envelope of fish embryos), polar body extrusion, and cytoplasmic streaming (fig. S4, A to C, and movie S1). Moreover, the micropyle, an opening in the chorion that serves as the sole entry point for sperm into zebrafish eggs, is present in $bncr^{-/-}$ eggs, and its size is similar to that of wild-type eggs (fig. S4D). These results suggest that Bouncer is not required for egg activation and micropyle formation.

Because eggs from $bncr^{-/-}$ females lack any apparent morphological defects yet do not develop beyond the one-cell stage, Bouncer might be required for fertilization and/or the initiation of early cleavage cycles. To distinguish between these possibilities, we first asked whether sperm can enter eggs lacking Bouncer. In vitro fertilization (IVF) of wild-type and Bouncer-deficient eggs with MitoTracker-labeled sperm allowed us to detect sperm only in wild-type eggs (50%) but never in Bouncer-deficient eggs (Fig. 2A), which suggests that Bouncer might play a role in sperm entry during fertilization. Consistent with the idea that Bouncer's sole function is to allow sperm to enter the egg, delivery of sperm into Bouncer-deficient eggs by intracytoplasmic sperm injection (ICSI) bypassed the requirement for Bouncer and restored embryonic development beyond the one-cell stage (Fig. 2B). Bouncer's key function is therefore in enabling sperm entry during fertilization.

[1]Research Institute of Molecular Pathology (IMP), Vienna Biocenter (VBC), 1030 Vienna, Austria. [2]Institute of Molecular Biotechnology of the Austrian Academy of Sciences (IMBA), Vienna Biocenter (VBC), 1030 Vienna, Austria.
*Corresponding author. Email: andrea.pauli@imp.ac.at

To gain further insight into Bouncer's function, we assessed its localization. Consistent with its predicted GPI anchorage, confocal imaging revealed that the fully functional, GFP-tagged Bouncer (Fig. 1E) localized to the egg membrane and to vesicles within the egg (Fig. 3A). Further, ubiquitous expression of a version of Bouncer lacking the C-terminal membrane anchor (GFP-Bncr^{noTM}) did not rescue the near-sterility of *bncr*$^{-/-}$ females (fig. S3, A and C), which suggests that membrane localization of Bouncer is required for its function.

The requirement for Bouncer at the egg membrane implies that it could promote the approach of sperm to the egg or sperm-egg binding/

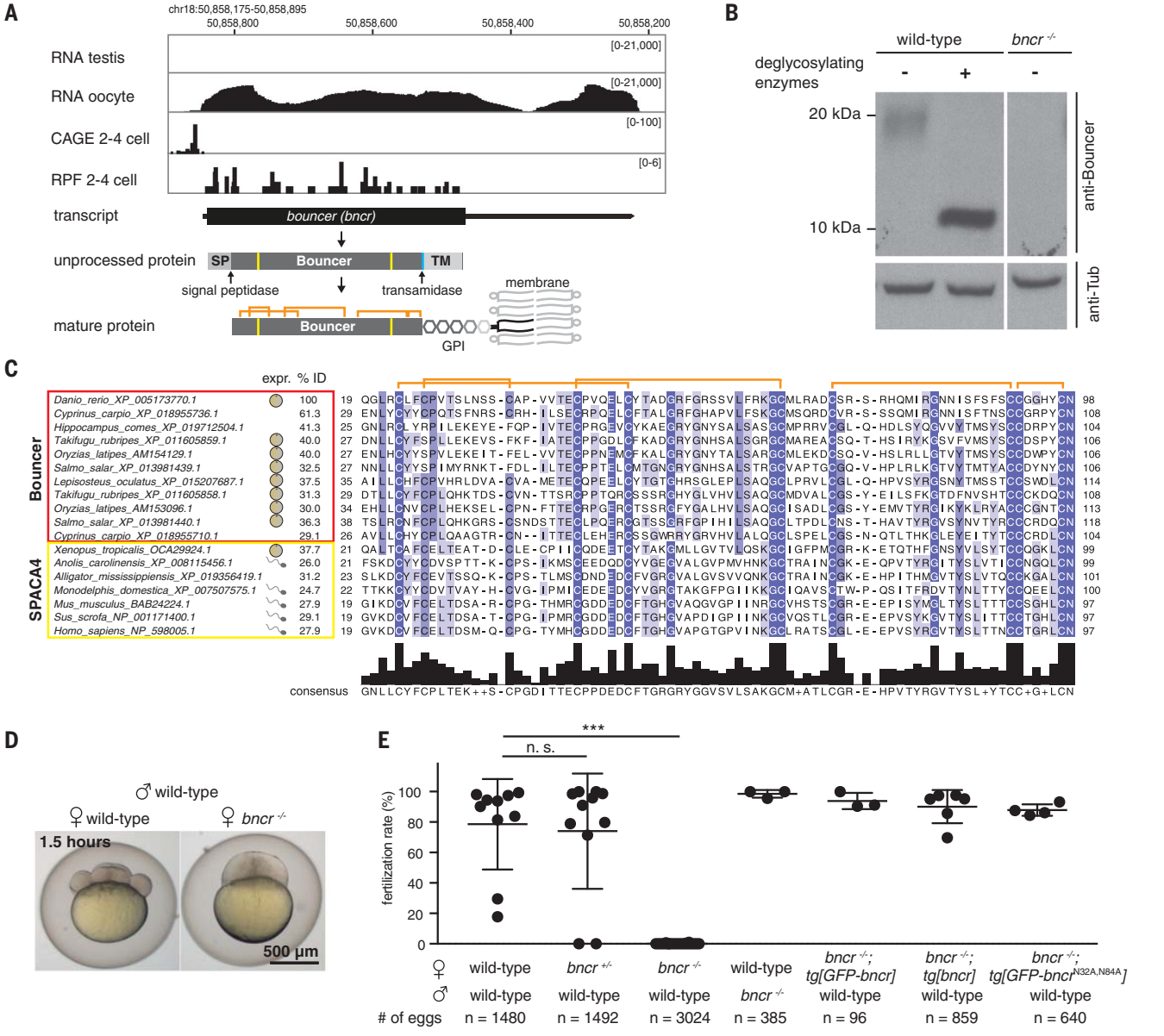


Fig. 1. Identification of Bouncer in fish. (**A**) Expression and genomic features of Bouncer. Coverage tracks for RNA sequencing, ribosome profiling (RPF) (*10*), and CAGE data (*12*) are shown. Genomic coordinates are based on GRCz10. SP, signal peptide; TM, transmembrane region; orange, predicted disulfide bonds; yellow, predicted N-glycosylation sites; turquoise, predicted transamidase cleavage site. (**B**) Endogenous Bouncer protein is glycosylated. Endogenous Bouncer is detected in the zebrafish egg by a Bouncer-specific antibody at a higher molecular weight than predicted (~20 kDa) but shifts down to the expected size (10 kDa) after treatment with deglycosylating enzymes. No Bouncer signal is detected in eggs from *bncr*$^{-/-}$ mutant females. (**C**) Protein sequence alignment of the mature domain of Bouncer/SPACA4 protein family members. Apart from the well-conserved cysteines (orange denotes predicted disulfide bonds), Bouncer/SPACA4 shows high amino acid divergence among different species (% ID, percent sequence identity to the mature domain of zebrafish Bouncer). The extent of the mature domain displayed here is based on the prediction for zebrafish Bouncer. For all species for which expression data were available [(*29*–*33*): human expression based on GTEx Portal; expressed sequence tags based on NCBI], *bouncer/Spaca4* RNA is restricted to either the male (symbol: sperm) or female (symbol: egg) germline. For sequences and accessions, see data S1 and S2 and table S1. Amino acid abbreviations: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr. (**D** and **E**) Lack of Bouncer causes near-sterility in female zebrafish. (D) Representative images of a developing, eight-cell stage embryo derived from a wild-type female, and an arrested, one-cell stage egg derived from a *bncr*$^{-/-}$ female 1.5 hours after mating. (E) Transgenically expressed, ubiquitin promoter–driven untagged, GFP-tagged, and nonglycosylatable Bouncer rescue the mutant phenotype. Data are means ± SD; *n* = number of eggs. ***$P < 0.0001$ (Kruskal-Wallis test with Dunn multiple-comparisons test); n.s., not significant.

fusion. Live cell imaging revealed that multiple MitoTracker-labeled sperm are recruited to the micropyle independently of Bouncer (Fig. 3B and movie S2). Thus, Bouncer does not provide an essential attractive cue that guides sperm toward the egg/micropyle.

During live imaging, multiple sperm entered the narrow opening of the micropyle simultaneously, rendering a more detailed analysis of sperm-egg binding capability infeasible. To investigate Bouncer's potential role in sperm-egg binding, we exposed the entire egg surface to sperm by removing the chorion. MitoTracker-labeled sperm remained bound to the surface of wild-type eggs in large clusters (>10 sperm) (Fig. 3, C and D). In contrast, only a few individual sperm (<10) remained attached to the majority of Bouncer-deficient eggs ($P < 0.002$) (Fig. 3, C and D). These results suggest that Bouncer promotes sperm-egg binding.

Bouncer shows a high degree of amino acid sequence divergence among different fish species, similar to other proteins involved in species specificity in mammals (fig. S5A). This raised the interesting possibility that Bouncer might contribute to the species specificity of fertilization in fish. To test this hypothesis, we generated *bncr*$^{-/-}$ zebrafish that ubiquitously express medaka Bouncer (*bncr*$^{-/-}$; tg[*ubi:medaka-bncr*]), henceforth called transgenic medaka Bouncer fish. Medaka was chosen because of its large evolutionary distance from zebrafish (~200 million years), its inability to cross-hybridize with zebrafish, and the low (40%) amino acid identity between the mature zebrafish and medaka Bouncer proteins (Fig. 4A). Expression of medaka Bouncer in *bncr*$^{-/-}$ females did not efficiently rescue fertility when crossed to wild-type male zebrafish (average fertilization rate of 0.45%, versus 0.11% for

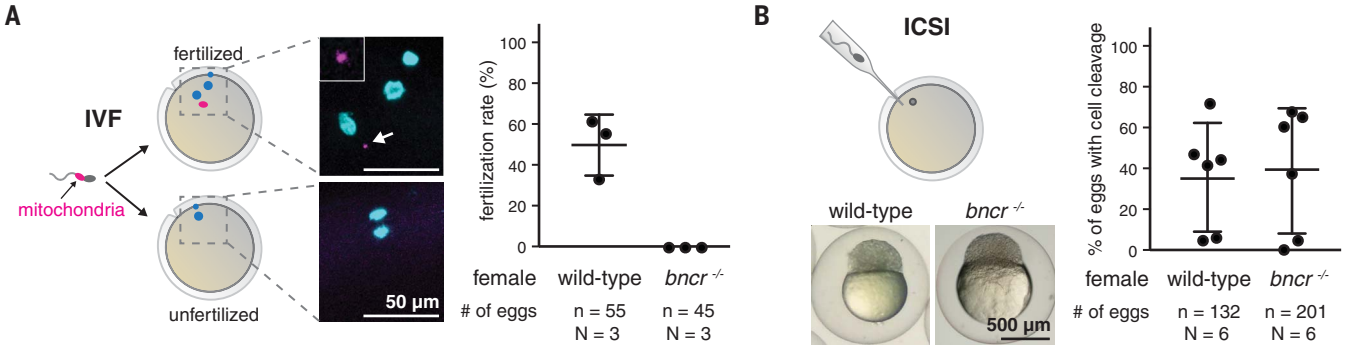


Fig. 2. Bouncer is required for sperm entry into the egg. (**A**) Sperm does not enter *bncr*$^{-/-}$ eggs. Left: Experimental setup. Wild-type sperm was stained with MitoTracker label and used for IVF of wild-type and *bncr*$^{-/-}$ eggs. Representative images are shown (arrow: MitoTracker signal, enlarged in white box). Right: Percentage of fertilized eggs, as indicated by the presence of one MitoTracker-labeled sperm and three DAPI signals (male nucleus, female nucleus, polar body). Means ± SD are indicated. (**B**) ICSI is able to rescue *bncr*$^{-/-}$ eggs. Top left: Experimental setup. Wild-type sperm was injected into wild-type or *bncr*$^{-/-}$ eggs. Cell cleavage was scored after 3 hours. Bottom left: Representative images. Right: Percentage of eggs that show cell cleavage. Means ± SD are indicated. *n* = total number of eggs; *N* = number of biological replicates.

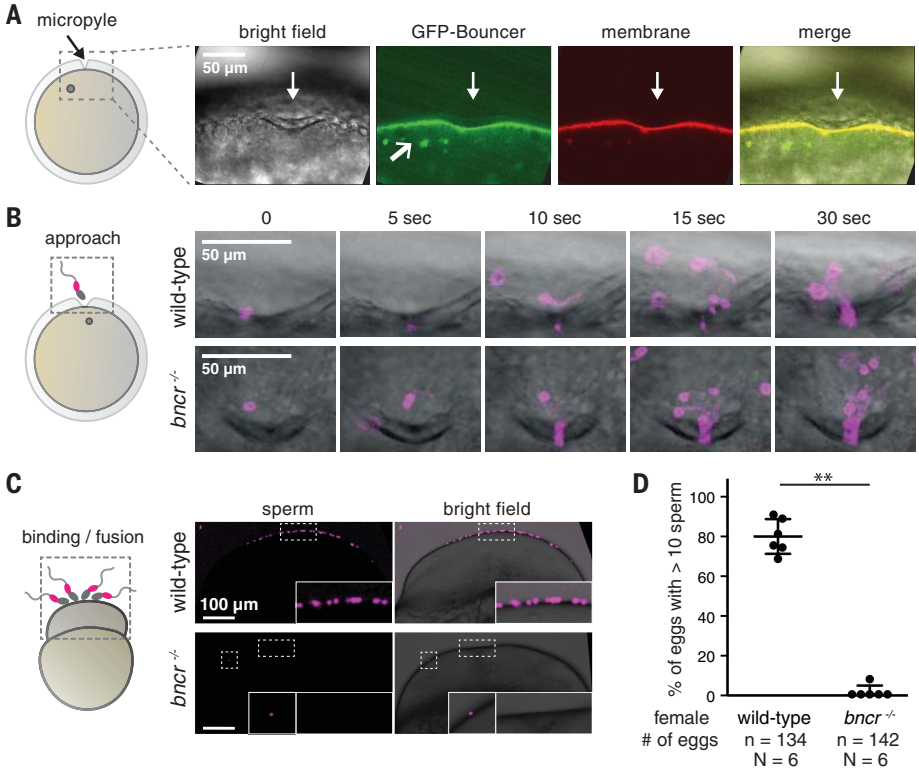


Fig. 3. Bouncer mediates binding between sperm and egg. (**A**) Bouncer localizes to the egg membrane and to vesicles. Confocal images of eggs expressing GFP-tagged Bouncer (green) and lyn-Tomato (membrane, red) during egg activation show that Bouncer localizes to the egg membrane around the micropyle (downward-pointing white arrow) and to vesicles (angled white arrow in GFP-Bouncer panel). Gray circle, egg nucleus. (**B**) Bouncer is not required for sperm approach. Left: Experimental setup. Right: Representative time series of multiple MitoTracker-labeled wild-type sperm (magenta) approaching the micropyle area of wild-type (top) and *bncr*$^{-/-}$ (bottom) eggs. (**C**) *bncr*$^{-/-}$ eggs are impaired in sperm-egg binding. Left: Experimental setup. Activated and dechorionated wild-type and *bncr*$^{-/-}$ eggs were incubated with MitoTracker-labeled wild-type sperm and gently washed. Right: Representative images of a wild-type egg (top, scored as >10) and a *bncr*$^{-/-}$ egg (bottom) with a single bound sperm. Boxed areas are also shown at higher magnification. (**D**) Quantification of sperm-egg binding. Eggs were classified as either >10 sperm bound or <10 sperm bound. Data are means ± SD (**$P < 0.002$, Mann-Whitney test; *n* = number of eggs; *N* = number of biological replicates).

bncr$^{-/-}$ and 78.6% for wild-type females) (Fig. 4B), supporting our hypothesis that Bouncer might influence species-specific gamete interaction. To directly test this possibility, we performed a series of IVF experiments. As expected, wild-type zebrafish eggs exhibited high fertilization rates with zebrafish sperm (average rate of 48.3%) but were not fertilized by medaka sperm (Fig. 4C). Moreover, zebrafish *bncr*$^{-/-}$ eggs were fertilized by neither sperm (Fig. 4C). Remarkably, eggs from transgenic medaka Bouncer females were fertilized by medaka, but not zebrafish, sperm (Fig. 4C), and eggs from females expressing both medaka Bouncer and zebrafish Bouncer could be fertilized by both sperm (fig. S5B). The average fertilization rate of all transgenic medaka Bouncer females (15) tested was 3.9% (fig. S5B). Whereas 6 of 15 females were infertile (average fertilization rate < 0.5%; fig. S5C), eggs from the remaining nine females were fertilized by medaka sperm at an average rate of 5.7% (Fig. 4C). Fertility rates of individual transgenic medaka Bouncer females were found to correlate with expression levels of medaka *bouncer* mRNA in eggs (fig. S5C), supporting a causal link between medaka Bouncer expression and medaka sperm entry. The resulting embryos were zebrafish-medaka hybrids and not haploid zebrafish embryos (fig. S5D). Hybrid embryos underwent cell cleavage and gastrulation (fig. S5E) and displayed anterior-posterior axis formation after 24 hours (Fig. 4D and fig. S5F) but did not survive past 48 hours. These results demonstrate that Bouncer is necessary and sufficient for mediating species-specific fertilization in fish.

Our finding that ectopic expression of another species' Bouncer is sufficient to allow cross-species fertilization strongly suggests that Bouncer has a direct, species-specific interaction partner on sperm. Additionally, the low average fertilization rate (5.7%) of medaka Bouncer–expressing zebrafish eggs by medaka sperm implies that other factors likely contribute to species-specific sperm-egg interaction. The identification of these factors and Bouncer's interaction partner on sperm will be crucial to unraveling the mechanism of species specificity of fertilization.

Thus far, the only known interacting membrane-bound proteins on vertebrate sperm and egg are IZUMO1 and JUNO in mammals (*1*–*3*, *7*). Whether these two proteins also play a role in mediating species-specific fertilization in vivo is, however, still unclear (*8*). In many organisms, species specificity of fertilization is mediated between proteins on the sperm membrane and those localized to the egg coat (*15*, *16*). For example, in sea urchin, the vitelline envelope protein EBR1 (egg bindin receptor protein–1) binds specifically to the sperm membrane protein bindin (*17*–*19*). Similarly, the egg coat protein VERL of abalone is species-specifically bound by lysin, a small secreted protein from sperm (*20*–*23*). Whereas lysin has no known homolog in vertebrates, VERL shows structural homology to the mammalian zona pellucida protein ZP2 (*22*), which was shown to be involved in species-specific binding of sperm to the zona pellucida in mouse and humans (*24*). Furthermore, from the side of the sperm, the mammalian sperm acrosomal protein zonadhesin binds species-specifically to the zona pellucida, even though it is not required for fertility (*25*). In contrast to these proteins, Bouncer mediates species-specific binding of sperm to the egg membrane, not to the egg coat.

Bouncer and SPACA4, both members of the large Ly6/uPAR superfamily, have opposing germline-specific expression patterns in externally versus internally fertilizing organisms.

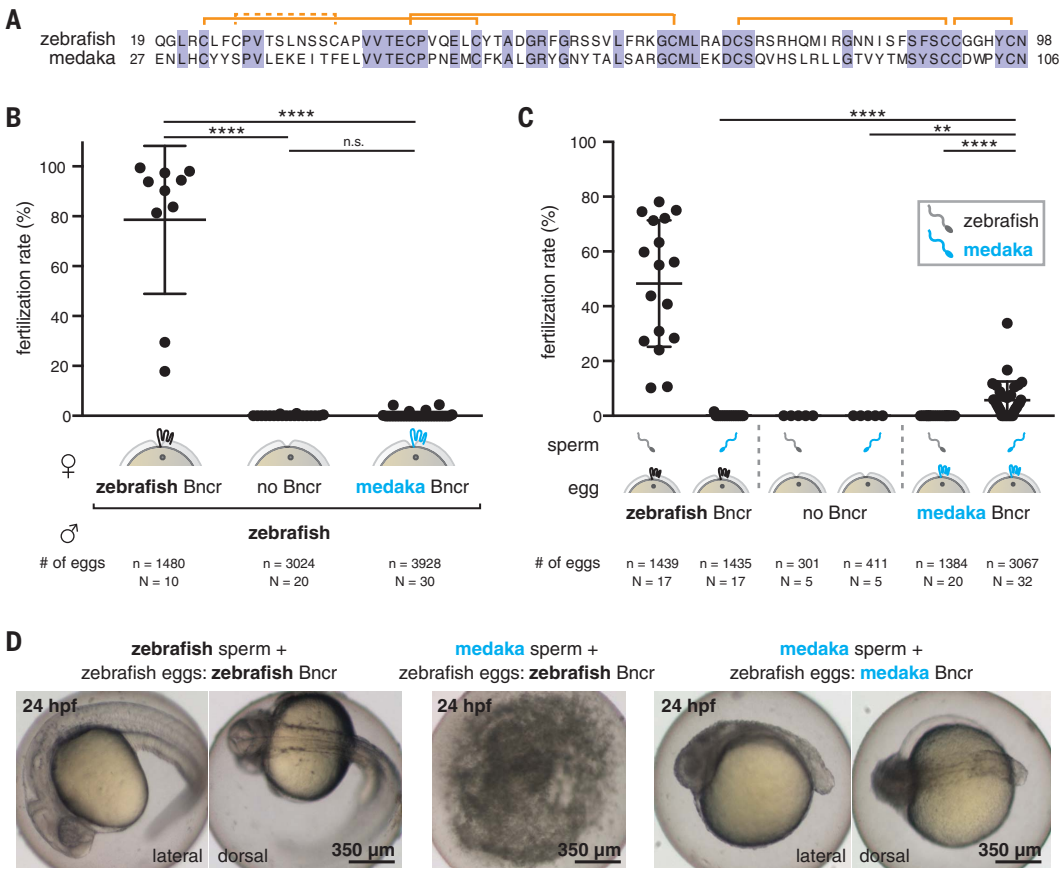


Fig. 4. Bouncer mediates species-specific fertilization. (**A**) Mature zebrafish and medaka Bouncer have only 40% amino acid sequence identity. Orange denotes predicted disulfide bonds (the dashed orange line denotes a disulfide bond predicted in zebrafish but not in medaka). (**B**) Medaka Bouncer does not efficiently rescue the fertilization defect of zebrafish *bncr*$^{-/-}$ females. Means ± SD are indicated [Kruskal-Wallis test with Dunn multiple-comparisons test: wild-type (wt) × wt versus *bncr*$^{-/-}$ × wt, adj. ****P < 0.0001; wt × wt versus medaka Bouncer × wt, adj. ****P < 0.0001; *bncr*$^{-/-}$ × wt versus medaka Bouncer × wt, n.s.; n = number of eggs; N = number of biological replicates]. (**C**) Medaka Bouncer is sufficient to allow entry of medaka sperm into zebrafish eggs. Medaka sperm did not fertilize wild-type zebrafish eggs, but medaka Bouncer–expressing zebrafish eggs had an average fertilization rate of 5.7% in IVF experiments. Data are shown for the subset of medaka Bouncer–expressing females (9 of 15) that were fertile (see fig. S5B for data of all 15 females tested). Data are means ± SD (Kruskal-Wallis test with Dunn multiple-comparisons test: medaka sperm on zebrafish versus medaka Bouncer eggs, adj. ****P < 0.0001; zebrafish sperm on medaka Bouncer eggs versus medaka sperm on medaka Bouncer eggs, adj. ****P < 0.0001; medaka sperm on *bncr*$^{-/-}$ zebrafish eggs versus medaka Bouncer eggs, adj. **P = 0.0052; n = number of eggs; N = number of biological replicates). (**D**) Fertilization of zebrafish eggs expressing only medaka Bouncer yields medaka-zebrafish hybrid embryos. Left: Wild-type zebrafish embryos fertilized by zebrafish sperm. Center: Wild-type zebrafish embryos are not fertilized by medaka sperm and decompose within 24 hours. Right: Zebrafish eggs expressing only medaka Bouncer are fertilized by medaka sperm and develop into hybrid embryos. hpf, hours post-fertilization.

The underlying reason for this is unclear, but one can speculate that in externally fertilizing species, oocyte expression of Bouncer contributes to postcopulatory female mate choice (also called cryptic female mate choice) (*26*). Vertebrates performing external fertilization cannot guarantee that only conspecific sperm reaches the egg by precopulatory mate choice (*27*, *28*). Oocyte-expressed proteins such as Bouncer could therefore support the selection of conspecific sperm. Our work on Bouncer also raises the intriguing possibility that SPACA4 might play an important role in mammalian fertilization, albeit from the side of the male. Although a knockout for murine *Spaca4* has not yet been reported, this idea is consistent with the localization of SPACA4 to the inner acrosomal membrane of sperm and the observed reduction of sperm-egg binding and fusion in vitro by incubation of sperm with antibody to SPACA4 (*14*). Future experiments that address the in vivo function of mammalian SPACA4 during fertilization will therefore be of interest. Given that both genes are restricted to the germline, our findings in fish may have direct relevance for fertilization in mammals.

REFERENCES AND NOTES

1. N. Inoue, M. Ikawa, A. Isotani, M. Okabe, *Nature* **434**, 234–238 (2005).
2. Y. Satouh, N. Inoue, M. Ikawa, M. Okabe, *J. Cell Sci.* **125**, 4985–4990 (2012).
3. E. Bianchi, B. Doe, D. Goulding, G. J. Wright, *Nature* **508**, 483–487 (2014).
4. K. Kaji *et al.*, *Nat. Genet.* **24**, 279–282 (2000).
5. K. Miyado *et al.*, *Science* **287**, 321–324 (2000).
6. F. Le Naour, E. Rubinstein, C. Jasmin, M. Prenant, C. Boucheix, *Science* **287**, 319–321 (2000).
7. K. Kato *et al.*, *Nat. Commun.* **7**, 12198 (2016).
8. E. Bianchi, G. J. Wright, *Philos. Trans. R. Soc. B* **370**, 20140101 (2015).
9. A. Pauli *et al.*, *Science* **343**, 1248636 (2014).
10. G.-L. Chew *et al.*, *Development* **140**, 2828–2834 (2013).
11. G.-L. Chew, A. Pauli, A. F. Schier, *Nat. Commun.* **7**, 11663 (2016).
12. V. Haberle *et al.*, *Nature* **507**, 381–385 (2014).
13. C. L. Loughner *et al.*, *Hum. Genomics* **10**, 10 (2016).
14. J. Shetty *et al.*, *J. Biol. Chem.* **278**, 30506–30515 (2003).
15. P. M. Wassarman, E. S. Litscher, *A Bespoke Coat for Eggs: Getting Ready for Fertilization* (Elsevier, ed. 1, 2016).
16. P. M. Wassarman, E. S. Litscher, *Int. J. Dev. Biol.* **52**, 665–676 (2008).
17. A. P. Stapper, P. Beerli, D. R. Levitan, *Mol. Biol. Evol.* **32**, 859–870 (2015).
18. N. Kamei, C. G. Glabe, *Genes Dev.* **17**, 2502–2507 (2003).
19. V. D. Vacquier, G. W. Moy, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 2456–2460 (1977).
20. J. D. Lyon, V. D. Vacquier, *Dev. Biol.* **214**, 151–159 (1999).
21. W. J. Swanson, V. D. Vacquier, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 6724–6729 (1997).
22. I. Raj *et al.*, *Cell* **169**, 1315–1326.e17 (2017).
23. C. A. Lewis, C. F. Talbot, V. D. Vacquier, *Dev. Biol.* **92**, 227–239 (1982).
24. M. A. Avella, B. Baibakov, J. Dean, *J. Cell Biol.* **205**, 801–809 (2014).
25. S. Tardif *et al.*, *J. Biol. Chem.* **285**, 24863–24870 (2010).
26. R. C. Firman, C. Gasparini, M. K. Manier, T. Pizzari, *Trends Ecol. Evol.* **32**, 368–382 (2017).
27. D. R. Levitan, *Am. Nat.* **191**, 88–105 (2018).
28. D. R. Levitan, D. L. Ferrell, *Science* **312**, 267–269 (2006).
29. B. Li *et al.*, *Sci. Rep.* **7**, 4200 (2017).
30. R. Marin *et al.*, *Genome Res.* **27**, 1974–1987 (2017).
31. J. Pasquier *et al.*, *BMC Genomics* **17**, 368 (2016).
32. A. M. Session *et al.*, *Nature* **538**, 336–343 (2016).
33. Z. Wang *et al.*, *Comp. Biochem. Physiol. D* **22**, 50–57 (2017).

ACKNOWLEDGMENTS

We thank K. Tessmar-Raible and B. M. Fontinha for generously providing medaka fish and expertise; A. Schier for his generous support during the start of this project and for valuable feedback on the manuscript; J. Gagnon for his help in generating the *bouncer* mutant and obtaining germline RNA-seq data; M. Novatchkova and L. E. Cabrera Quio for help with RNA-seq data mapping and gene expression analyses; K. Panser for help with genotyping; the IMP animal facility personnel, especially J. König and F. Ecker, for their excellent care of our fish; P. Pasierbek from the BioOptics core facility for support in microscopy; T. Heuser and N. Fellner from the VBCF EM facility for help with EM; M. Madalinski for synthesizing Bouncer peptides for antibody production; J. Farrell for providing the sfGFP plasmid; C.-P. Heisenberg for providing the tg[*lyn-tdTomato*] fish line; the Pauli lab for discussions; the unidentified scientist at the 20th Anniversary Symposium of the Zebrafish Course at the Marine Biological Laboratory in Woods Hole, MA, for suggesting the name "Bouncer" to us; and A. Anderson (Life Science Editors), A. Stark, C.-P. Heisenberg, L. Cochella, E. Tanaka, M. Ikawa, Y. Fujihara, and B. Podbilewicz for helpful comments on the manuscript. **Funding:** Supported by the Research Institute of Molecular Pathology (IMP), Boehringer Ingelheim, and the Austrian Academy of Sciences; a DOC Fellowship from the Austrian Academy of Sciences (S.H.); and HFSP Career Development Award CDA00066/2015 and the FWF START program (A.P.). **Author contributions:** S.H. and A.P. conceived the study; S.H. performed most experiments except experiments regarding species specificity, which were performed by K.R.G., and generation of RNA-seq data and *bncr*$^{-/-}$ fish, which were performed by A.P.; S.H., K.R.G., and A.P. analyzed the data; A.S. and A.P. performed the phylogenetic analysis; and S.H. and A.P. wrote the manuscript with input from K.R.G. and A.S. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** RNA-seq data first reported here were deposited at the Gene Expression Omnibus (GEO) and are available under GEO acquisition number GSE111882. All other data are available in the manuscript or the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6406/1029/suppl/DC1
Materials and Methods
Figs. S1 to S5
Table S1
Movies S1 and S2
Data S1 and S2
References (*34*–*57*)

27 March 2018; accepted 11 July 2018
10.1126/science.aat7113

CANCER

Minimal functional driver gene heterogeneity among untreated metastases

Johannes G. Reiter[1,2]*†, Alvin P. Makohon-Moore[3]†, Jeffrey M. Gerold[2]†, Alexander Heyde[2], Marc A. Attiyeh[3], Zachary A. Kohutek[4], Collin J. Tokheim[5], Alexia Brown[3], Rayne M. DeBlasio[3], Juliana Niyazov[3], Amanda Zucker[3], Rachel Karchin[5,6], Kenneth W. Kinzler[7,8,9], Christine A. Iacobuzio-Donahue[3,10], Bert Vogelstein[7,8,9,11], Martin A. Nowak[2,12]*

Metastases are responsible for the majority of cancer-related deaths. Although genomic heterogeneity within primary tumors is associated with relapse, heterogeneity among treatment-naïve metastases has not been comprehensively assessed. We analyzed sequencing data for 76 untreated metastases from 20 patients and inferred cancer phylogenies for breast, colorectal, endometrial, gastric, lung, melanoma, pancreatic, and prostate cancers. We found that within individual patients, a large majority of driver gene mutations are common to all metastases. Further analysis revealed that the driver gene mutations that were not shared by all metastases are unlikely to have functional consequences. A mathematical model of tumor evolution and metastasis formation provides an explanation for the observed driver gene homogeneity. Thus, single biopsies capture most of the functionally important mutations in metastases and therefore provide essential information for therapeutic decision-making.

The clonal evolution model of cancer proposes that cells accrue advantageous mutations and clonally expand so that these mutations are eventually present in all tumor cells (*1*–*4*). Recent studies reported mutations in putative driver genes that were only present in subpopulations of tumor cells (*5*, *6*). The extent to which the acquisition of advantageous mutations continues after the initiation of the primary tumor (*7*) or during metastasis formation is unknown (*8*, *9*). The growing list of putative driver genes and the increased sensitivity of next-generation sequencing have facilitated the discovery of subclonal driver gene mutations within a tumor (*5*, *10*). Nevertheless, the evolutionary dynamics and the clinical importance of driver gene mutation heterogeneity in solid tumors are not fully understood.

Cells acquire a few mutations during each division because of imperfect DNA replication; hence, any population of cells is genetically heterogeneous (*11*). Because cancer cells continue to divide after cancer initiation, many new mutations are expected to be present in tumor subpopulations. However, to assess functional heterogeneity, advantageous mutations in putative driver genes must be distinguished from neutral replication errors in those genes. For example, within oncogenes, only few recurrently mutated positions are functional, and therefore, many mutations—even in driver genes—may not have important functional consequences. Moreover, although metastatic disease is responsible for most cancer-related deaths, the heterogeneity of driver gene mutations has predominantly been evaluated in primary tumors. Biopsies of metastatic lesions are not readily available and typically are acquired after exposure to toxic and mutagenic chemotherapies. These treatments can induce selective bottlenecks and confound the interpretation of genetic alterations.

Because driver gene mutations increasingly inform clinical treatment decisions, undetected driver heterogeneity among metastases poses a barrier to the success of this precision medicine approach (*12*). If the founding cells of different metastases carry distinct driver gene mutations, disease progression and treatment could be fundamentally more complex than expected from a primary tumor biopsy alone. Additional driver gene mutations might be present in all or in a subset of metastases (Fig. 1). In both scenarios, more biopsies would be necessary for accurate diagnosis and optimal treatment. Here, we comprehensively analyzed the evidence for driver gene mutation heterogeneity among untreated metastases across cancer types. We also developed a mathematical model to determine the evolutionary mechanisms that give rise to intermetastatic driver mutation heterogeneity.

We analyzed data from 20 cancer patients for whom genome- or exome-wide sequencing was performed for at least two distinct treatment-naïve metastases (*13*–*19*). In total, we studied 115 samples, including 76 untreated metastases samples from diverse tissues (mean of 3.8 and median of 3 metastases per patient) (fig. S1 and table S1). We assessed somatic mutations of patients with pancreatic, endometrial, colorectal, breast, gastric, lung, melanoma, and prostate cancer (Fig. 2A). We classified nonsynonymous variants into putative driver and passengers mutations according to the The Cancer Genome Atlas consensus list of 299 putative driver genes (*10*). To allow for a consistent interpretation of driver gene mutation heterogeneity, we excluded two hypermutated subjects with more than 1000 nonsynonymous mutations and focused on the remaining 18 subjects. In these subjects, we found a median of 4.5 mutated driver genes (range 2 to 18) (Fig. 2A).

To determine the evolutionary timing of somatic mutations, we inferred cancer phylogenies and mapped all variants onto evolutionary trees (supplementary materials, materials and methods, and fig. S2) (*20*). We classified mutations into those present in all metastases (MetTrunk; hereafter referred to as "trunk") and those present in a subset of metastases (MetBranch; hereafter referred to as "branch") (Fig. 2B). We observed similar numbers of nonsynonymous or splice-site variants (hereafter referred to as nonsynonymous) in both categories (Fig. 2A). By contrast, trunks exhibited a twofold enrichment of the ratio of driver gene mutations to nonsynonymous mutations compared with branches (9.1 versus 4.0%; two-sided paired t test, $P = 0.004$) (Fig. 3A). Nevertheless, we observed mutations in driver genes that were heterogeneous among metastases for 12 of 18 subjects.

To investigate whether heterogeneous mutations in putative driver genes were likely to be functional, we used a variety of approaches. We found that a large proportion of nonsynonymous variants in driver genes along trunks were previously detected at least once in other cancers [Catalogue Of Somatic Mutations In Cancer (COSMIC); 37.8%, 31 of 82], whereas a much smaller proportion along branches was present in COSMIC (15.6%, 5 of 32; two-sided Fisher's exact test, $P = 0.025$) (Fig. 3B). The fraction of driver gene mutations in branches in COSMIC was in fact similar to that of passenger gene mutations in either trunks or branches (14.1%, 128 of 905, and 12.5%, 89 of 712). Because mutations that are true drivers are often recurrent, we investigated how frequently identical

[1]Canary Center for Cancer Early Detection, Department of Radiology, Stanford University School of Medicine, Palo Alto, CA 94305, USA. [2]Program for Evolutionary Dynamics, Harvard University, Cambridge, MA 02138, USA. [3]The David M. Rubenstein Center for Pancreatic Cancer Research, Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA. [4]Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA. [5]Department of Biomedical Engineering, Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD 21218, USA. [6]Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA. [7]The Ludwig Center, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA. [8]The Sol Goldman Pancreatic Cancer Research Center, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA. [9]Sidney Kimmel Cancer Center, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA. [10]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA. [11]Howard Hughes Medical Institute, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA. [12]Department of Organismic and Evolutionary Biology and Department of Mathematics, Harvard University, Cambridge, MA 02138, USA.
***Corresponding author. Email: johannes.reiter@stanford.edu (J.G.R.); martin_nowak@harvard.edu (M.A.N.)** †These authors contributed equally to this work.

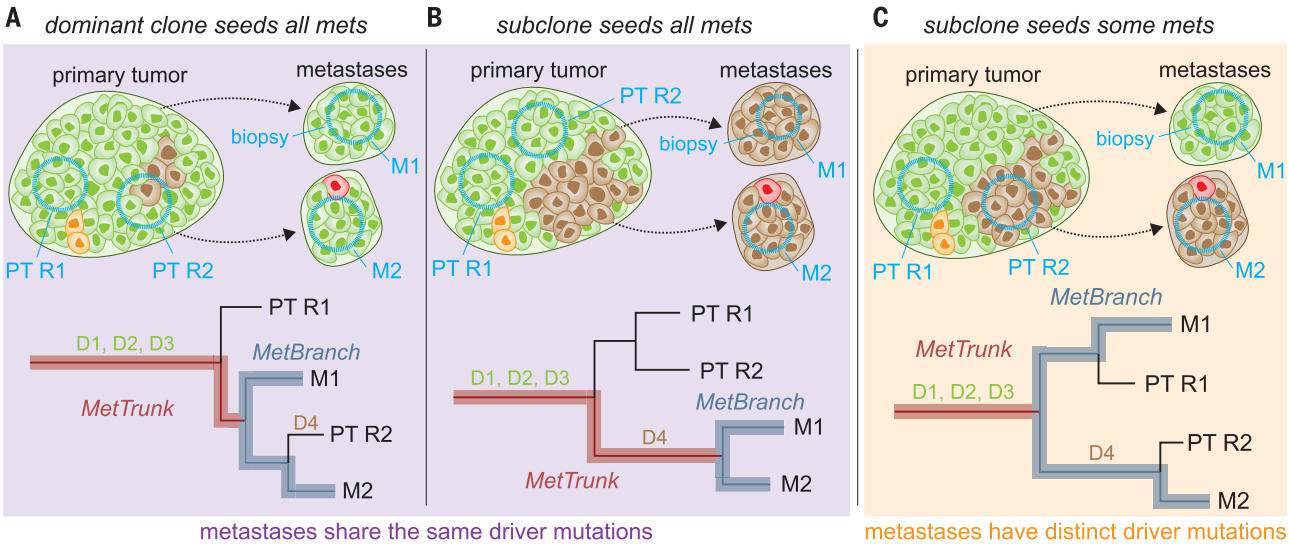


Fig. 1. Three scenarios of heterogeneity of mutations in driver genes. The original clone (green cells) contains three driver gene mutations (D1, D2, and D3). Brown, yellow, and red cells acquired additional driver mutations during the growth of the primary tumor (PT) and may expand to form detectable subpopulations (brown) that can seed metastases. (Top) Seeding subpopulations and biopsies (blue circles) of different regions (R1 and R2) of the PT and of distinct metastases (M1 and M2). (Bottom) Reconstructed cancer phylogenies from those biopsies. (**A**) Original clone seeds all metastases. All metastases share same founding driver mutations. Subclones with additional driver mutations (D4) evolve too late to seed metastases but might be detectable in the PT. (**B**) A single highly metastatic subclone evolves and gives rise to all metastases. All metastases share same founding driver mutations. (**C**) A new subclone with an additional driver mutation (D4) evolves and independently seeds metastases. PT regions and metastases exhibit driver mutation heterogeneity.

nonsynonymous variants were found in COSMIC. Whereas variants in driver genes along trunks on average occurred in 0.32% COSMIC samples (occurrence mean of 82.0 in 25,516 COSMIC samples), driver gene mutations acquired along branches occurred more than 100-fold less frequently (0.0016%; Wilcoxon rank-sum test, P = 0.008) (Fig. 3C).

We then used several methods to predict the functional impact of 1755 nonsynonymous variants along trunks and branches. We found that driver gene mutations acquired along trunks were more likely to have predicted functional consequences (Fig. 3, D to F, and fig. S3). Variants with the most likely protein-changing effects (mutation consequences with high impact, such as frameshift or nonsense mutations) were frequently observed in driver genes along trunks but rarely observed along branches (30.5 versus 6.3%; two-sided Fisher's exact test, P = 0.006) (Fig. 3D). The frequency of high-impact variants in driver genes along branches was no higher than that in passenger genes. FATHMM (*21*) predicted significantly stronger functional effects for driver gene mutations along trunks than along branches (mean scores of −2.1 versus 1.0; scores below −0.75 indicate likely driver mutation; two-sided Wilcoxon rank-sum test, P < 0.001) (Fig. 3E). Similarly, CHASMplus (*22*) predicted significantly higher gene-weighted scores for driver gene mutations along trunks than along branches (mean scores 0.47 versus 0.16; higher values indicate likely functional effects; two-sided Wilcoxon rank-sum test, P < 0.001) (Fig. 3F).

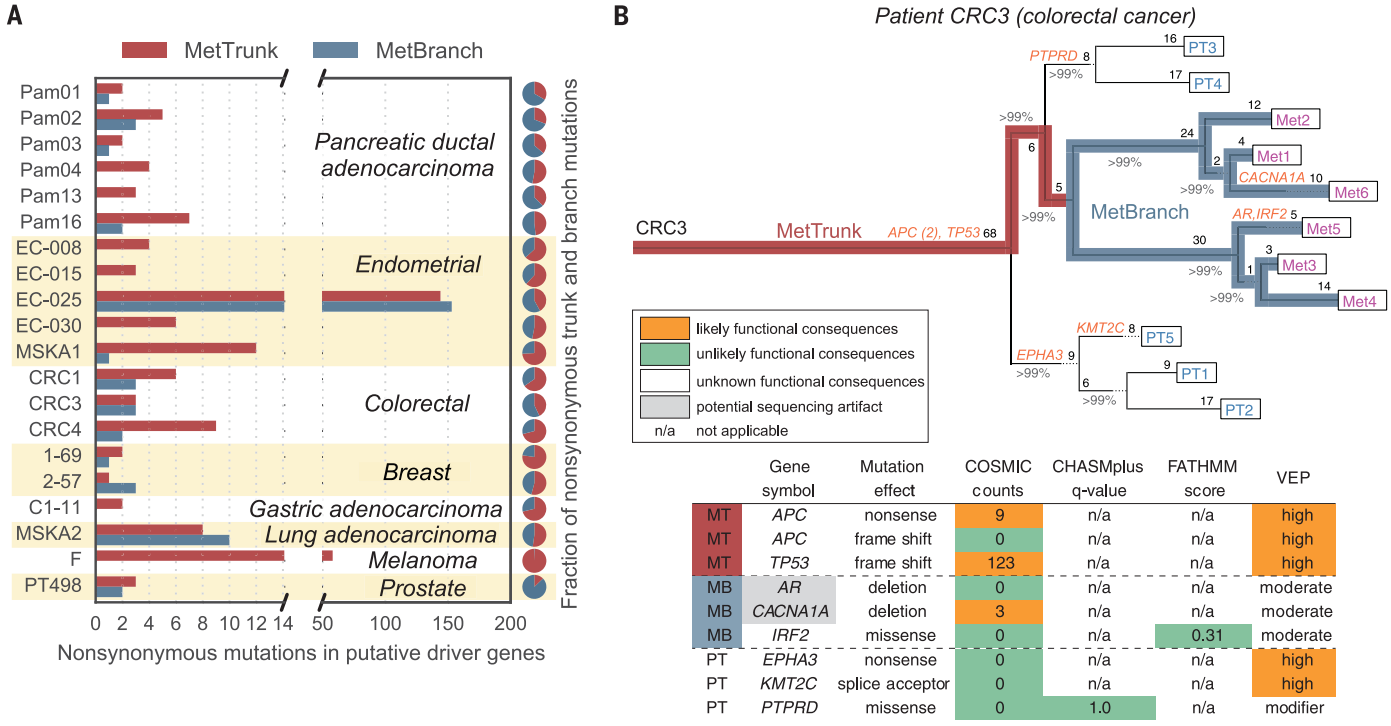
To identify the evolutionary determinants of intermetastatic heterogeneity, we developed a mathematical framework in order to assess how rates of growth, mutation, and dissemination give rise to driver gene mutation heterogeneity (supplementary text) (*23*, *24*). The original clone in the primary tumor grows with a rate of $r_0 = b_0 - d_0$ per day (birth rate is b_i and death rate is d_i for each clone i) and disseminates cells to distant sites with rate q_0 per day (Fig. 4A). When a cell divides, a daughter cell can acquire an additional driver mutation with probability u. This model produces intermetastatic heterogeneity if not all detectable metastases were seeded from the same subclone in the primary tumor.

Following previously measured growth and selection parameters, we assume a growth rate of r_0 = 1.24% per day and a relative growth advantage of a driver gene mutation of s = 0.4% ($s = b_i/b_0 - 1$) (*25*, *26*). To mimic the composition of our cohort, we considered the first four metastases that reach a detectable size of 10^8 cells (~1 cm^3). We found that the probability of intermetastatic driver heterogeneity is 10.5% (d = 0.2475, $q = 10^{-7}$) (Fig. 4). The original founding clone of the primary tumor most likely seeds all detectable metastases (Fig. 1A, green cells). The increased growth rate conferred by a new driver mutation is insufficient to compensate for the time spent waiting for the driver mutation to occur (figs. S4 and S5).

The model reveals that the probability of observing intermetastatic driver heterogeneity increases when the primary tumor grows very slowly before metastases are seeded, the average growth advantage of additional driver mutations is very large, and the driver gene mutation rate is high (fig. S6C). By contrast, a high dissemination rate produces less intermetastatic heterogeneity because metastases are established before driver subclones greatly expand (Fig. 4E and fig. S7C). For very high driver growth advantages but slowly growing cancers, another scenario is possible: that all metastases are seeded from the same highly advantageous subclone (Fig. 1B). Last, if driver mutations instead increase the dissemination rate, an almost 10-fold increase is required to produce intermetastatic driver heterogeneity (Fig. 4F and fig. S8).

In real patients, we expect less intermetastatic heterogeneity for several reasons. First, driver gene mutations may not confer the same advantage in the microenvironment of the primary tumor and of a distant site, reducing the probability of heterogeneity (fig. S9). Second, primary tumor growth may slow down because of space or nutrient constraints or surgical removal, also reducing the expected intermetastatic heterogeneity (fig. S10). Third, advanced cancer cells have already acquired multiple driver gene mutations in various pathways, possibly reducing the number of additionally available driver gene mutations that confer a substantial selective advantage (fig. S6B).

Overall, we observed a depletion of heterogeneous mutations in putative driver genes among metastases (Fig. 3). Moreover, the majority of those that were observed had only weak or no predicted functional effects. These results are compatible with multiple recent studies on neutrally evolving cancers after transformation (*7*, *27*, *28*). However, the mathematical framework demonstrates that a lack of intermetastatic driver heterogeneity does not imply neutral evolution but can also be explained by various other factors,



	Gene symbol	Mutation effect	COSMIC counts	CHASMplus q-value	FATHMM score	VEP
MT	*APC*	nonsense	9	n/a	n/a	high
MT	*APC*	frame shift	0	n/a	n/a	high
MT	*TP53*	frame shift	123	n/a	n/a	high
MB	*AR*	deletion	0	n/a	n/a	moderate
MB	*CACNA1A*	deletion	3	n/a	n/a	moderate
MB	*IRF2*	missense	0	n/a	0.31	moderate
PT	*EPHA3*	nonsense	0	n/a	n/a	high
PT	*KMT2C*	splice acceptor	0	n/a	n/a	high
PT	*PTPRD*	missense	0	1.0	n/a	modifier

Fig. 2. Most mutations in putative driver genes occur on the trunk of metastases. (**A**) Twenty patients with 76 untreated metastases. Thirteen patients acquired mutations in putative driver genes along the MetBranch (MB), whereas seven did not. (**B**) Inferred phylogeny of a colorectal cancer exhibits intermetastatic driver mutation heterogeneity. Nonsynonymous mutations in driver genes are denoted in orange. Percentages denote branch confidence. Integers denote number of point mutations per branch. Table shows predicted functional effects of mutations in driver genes. Heterogeneous driver mutations were predicted to have no functional effect or were likely sequencing artifacts [low coverage and low variant allele frequency (VAF) across all sites]. MetTrunk (MT) denotes that variant was acquired on the trunk of all metastases. Sample origins, rectum, PT1-5; liver, Met1-6.

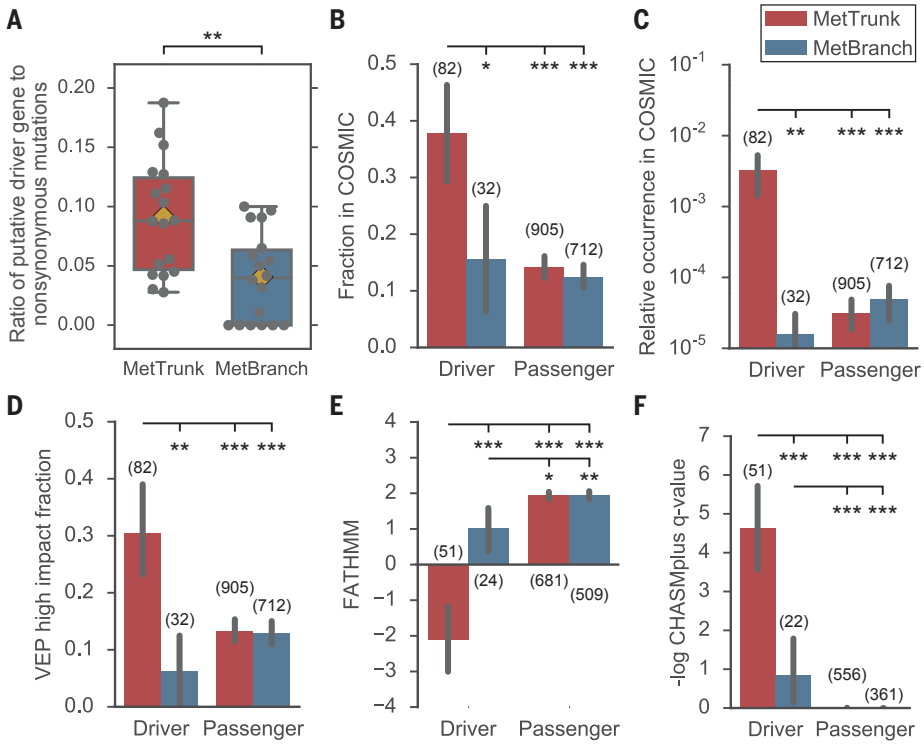


Fig. 3. Predicted functional mutations in putative driver genes are strongly enriched along metastases trunks. (**A**) Ratio of driver gene mutations to nonsynonymous mutations is enriched by twofold along trunks compared with branches. Orange diamond denotes mean, and black bar denotes median (two-sided paired *t* test, $P = 0.004$). (**B**) Fraction of nonsynonymous variants in driver genes along MetTrunk in COSMIC was 38% compared with 16% along MetBranch (two-sided Fisher's exact test, $P = 0.025$). (**C**) Relative occurrence of variants in driver genes along MetTrunk in individual COSMIC samples was 0.32% compared with 0.0016% along MetBranch (two-sided Wilcoxon rank-sum test, $P = 0.008$). (**D**) Variant Effect Predictor (VEP) inferred that 30 and 6% of driver gene mutations were of high impact along MetTrunk and MetBranch, respectively (two-sided Fisher's exact test, $P = 0.006$). (**E** and **F**) FATHMM (value below −0.75 indicates likely driver mutation) and CHASMplus predicted increased functional consequences for variants in driver genes in MetTrunk. Two-sided Wilcoxon rank-sum tests were used. Thick black bars denote 90% confidence interval. No other statistically significant differences were observed. Numbers in brackets denote number of variants in each group. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

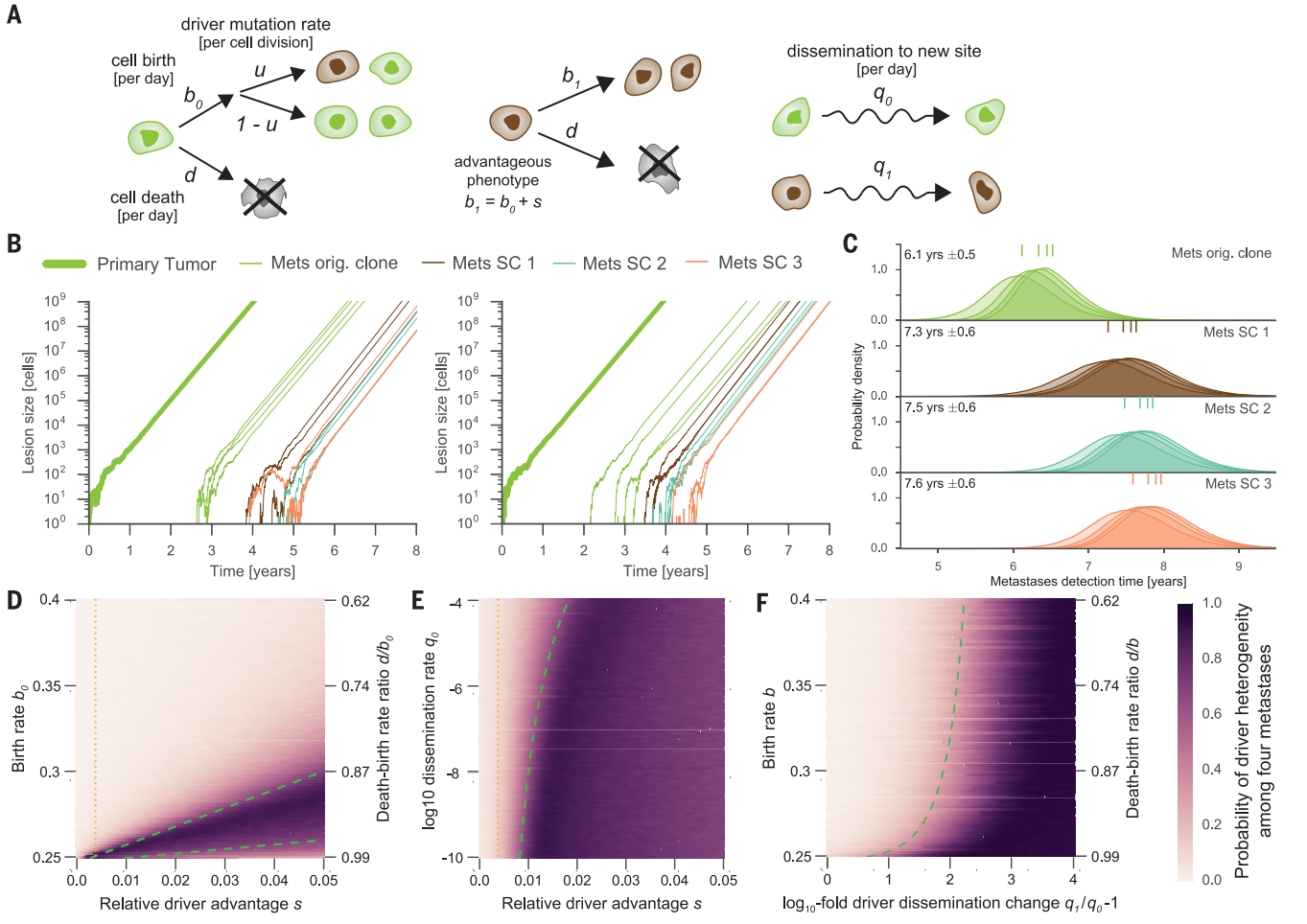


Fig. 4. Mathematical analysis provides an explanation for intermetastatic driver gene mutation homogeneity or heterogeneity. (**A**) Primary tumor expands stochastically from a single advanced cancer cell and seeds metastases. Cells of original clone (green) divide at rate b_0 and die at rate d per day. Additional driver mutations increase the birth rate to $b_1 = b_0(1 + s)$, where s denotes the relative driver advantage [$b_1 \geq b_0$, $q_0 = q_1$; (B) to (E)], or increase the dissemination rate [$q_1 \geq q_0$, $b_1 = b_0$ (F)]. (**B**) Representative model realizations for typical parameter values. Growth rate $r_0 = 1.24\%$ per day, $s = 0.4\%$, and dissemination rate $q_0 = 10^{-7}$ per cell per day. (**C**) Distribution of metastases detection times for parameter values in (B). Numbers denote mean ± SD. Colored marks show mean detection times of first, second, third, and fourth metastases seeded by the corresponding subclone (SC). (**D** to **F**) Probability of distinct driver mutations among four metastases. Green dashed lines depict bounds separating parameter regions of likely intermetastatic driver homogeneity from heterogeneity. Orange dotted lines denote $s = 0.4\%$. (D) Fixed $q_0 = 10^{-7}$. (E) Fixed death-birth rate ratio $d/b_0 = 0.95$. (F) Fixed $q_0 = 10^{-7}$. Other parameter values are $d = 0.2475$ and driver mutation rate $u = 3.4 \times 10^{-5}$ per cell division.

including primary tumor growth dynamics (Fig. 4). Furthermore, growth rates may saturate and fitness gains of additional driver gene mutations become smaller because available resources (such as nutrients and oxygen) are already almost optimally utilized—a phenomenon that is observed in bacterial evolution (*29*).

Several limitations of this study should be noted. First, we exclusively focused on single-nucleotide variants and small insertions and deletions because their functionality can be predicted with multiple methods, and their heterogeneity has immediate clinical consequences for therapy selection (*12*). We did not assess recurrent noncoding, copy-number, or epigenetic alterations because functional prediction methods for them are not yet available. Second, we cannot exclude the possibility that mutations in yet-undiscovered driver genes of metastases are heterogeneous. Third, we could not evaluate micrometastases that are not visible clinically.

Because therapy selection and treatment success of previously untreated patients increasingly depends on the identification of genetic alterations, it will be critical to extend this analysis to larger cohorts and more cancer types in order to investigate whether minimal driver gene mutation heterogeneity is a general phenomenon of advanced disease. This pan-cancer analysis of untreated metastases suggests that a single biopsy accurately represents the driver gene mutations of a patient's metastases.

REFERENCES AND NOTES

1. P. C. Nowell, *Science* **194**, 23–28 (1976).
2. E. R. Fearon, B. Vogelstein, *Cell* **61**, 759–767 (1990).
3. R. H. Hruban, M. Goggins, J. Parsons, S. E. Kern, *Clin. Cancer Res.* **6**, 2969–2972 (2000).
4. M. Greaves, C. C. Maley, *Nature* **481**, 306–313 (2012).
5. M. Jamal-Hanjani *et al.*, *N. Engl. J. Med.* **376**, 2109–2121 (2017).
6. M. Gerlinger *et al.*, *N. Engl. J. Med.* **366**, 883–892 (2012).
7. A. Sottoriva *et al.*, *Nat. Genet.* **47**, 209–216 (2015).
8. S. Turajlic, C. Swanton, *Science* **352**, 169–175 (2016).
9. J. Massagué, A. C. Obenauf, *Nature* **529**, 298–306 (2016).
10. M. H. Bailey *et al.*, *Cell* **173**, 371–385.e18 (2018).
11. F. Blokzijl *et al.*, *Nature* **538**, 260–264 (2016).
12. B. Vogelstein *et al.*, *Science* **339**, 1546–1558 (2013).
13. J. Z. Sanborn *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 10995–11000 (2015).
14. M. K. H. Hong *et al.*, *Nat. Commun.* **6**, 6605 (2015).
15. T.-M. Kim *et al.*, *Clin. Cancer Res.* **21**, 4461–4472 (2015).
16. W. J. Gibson *et al.*, *Nat. Genet.* **48**, 848–855 (2016).
17. A. P. Makohon-Moore *et al.*, *Nat. Genet.* **49**, 358–366 (2017).
18. D. Brown *et al.*, *Nat. Commun.* **8**, 14944 (2017).
19. E. Pectasides *et al.*, *Cancer Discov.* **8**, 37–48 (2018).
20. J. G. Reiter *et al.*, *Nat. Commun.* **8**, 14114 (2017).
21. H. A. Shihab, J. Gough, D. N. Cooper, I. N. M. Day, T. R. Gaunt, *Bioinformatics* **29**, 1504–1510 (2013).
22. C. Tokheim, R. Karchin, *bioRxiv* 10.1101/313296 (2018).
23. H. Haeno *et al.*, *Cell* **148**, 362–375 (2012).
24. R. Durrett, *Branching Process Models of Cancer* (Springer, 2015).

25. I. Bozic *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **107**, 18545–18550 (2010).
26. H. Furukawa, R. Iwata, N. Moriyama, *Pancreas* **22**, 366–369 (2001).
27. I. Bozic, J. M. Gerold, M. A. Nowak, *PLOS Comput. Biol.* **12**, e1004731 (2016).
28. M. J. Williams *et al*., *Nat. Genet.* **50**, 895–903 (2018).
29. B. H. Good, M. J. McDonald, J. E. Barrick, R. E. Lenski, M. M. Desai, *Nature* **551**, 45–50 (2017).

ACKNOWLEDGMENTS

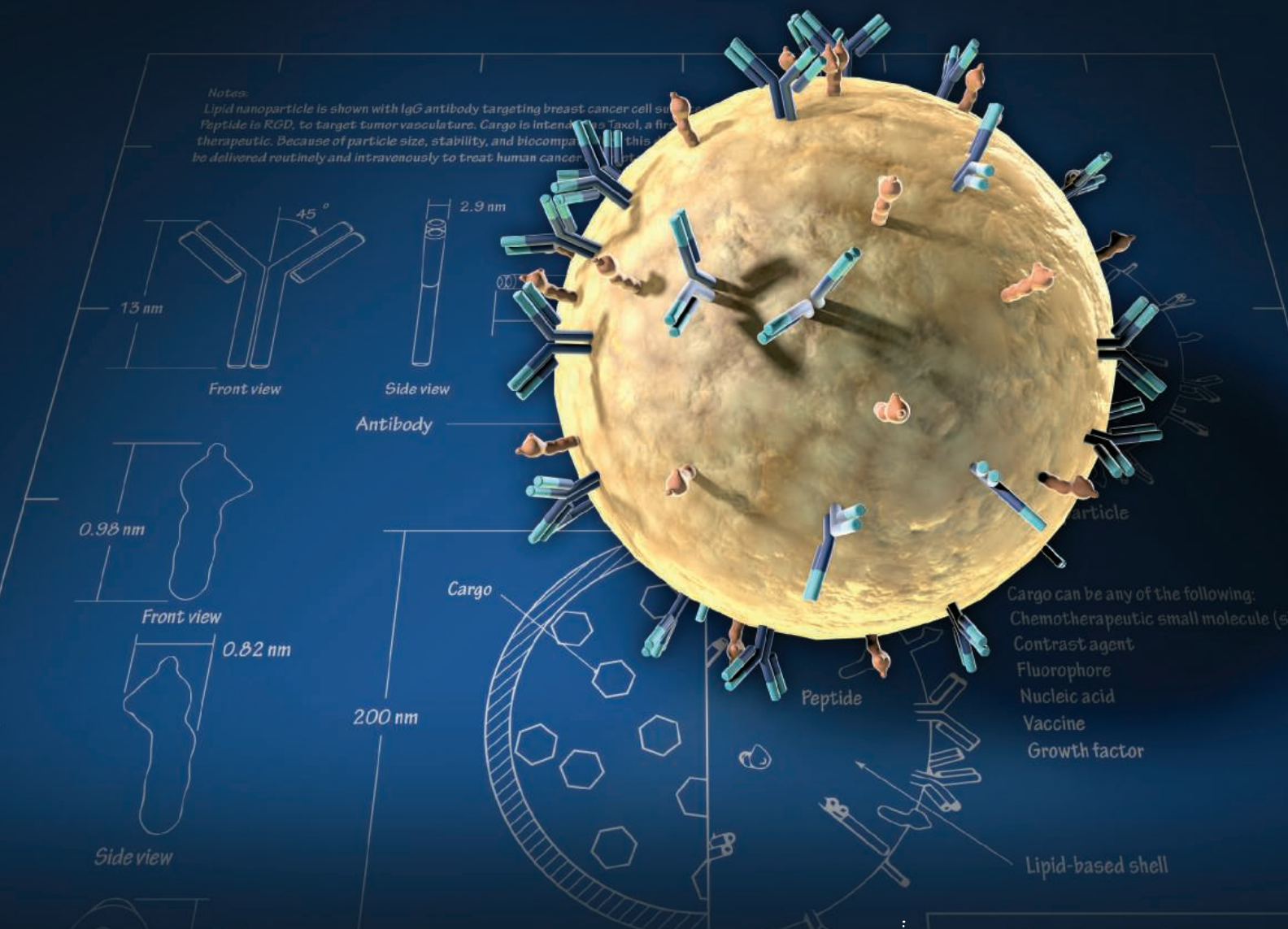
Funding: This work was supported by the National Institutes of Health grants K99CA229991 (J.G.R.), CA179991 (C.A.I.-D.), F31CA180682 (A.P.M.-M.), T32 CA160001-06 (A.P.M.-M.), F31CA200266 (C.J.T.), U24CA204817 (R.K.), CA43460 (B.V.), as well as by the Lustgarten Foundation for Pancreatic Cancer Research, The Sol Goldman Center for Pancreatic Cancer Research, The Virginia and D. K. Ludwig Fund for Cancer Research, an Erwin Schrödinger fellowship (J.G.R.; Austrian Science Fund FWF J-3996), a Landry Cancer Biology fellowship (J.M.G.), and the Office of Naval Research grant N00014-16-1-2914. **Author contributions:** J.G.R., A.P.M.-M., C.A.I.-D., B.V., and M.A.N. conceived and designed the study. A.P.M.-M., M.A.A., Z.A.K., A.B., R.M.D., J.N., A.Z., and C.A.I.-D. performed autopsies. A.P.M.-M., M.A.A., Z.A.K., K.W.K., C.A.I.-D., and B.V. generated sequencing data. J.G.R. performed computational analysis. J.G.R., J.M.G., A.H., and M.A.N. performed mathematical modeling. C.J.T. and R.K. performed CHASMplus analysis. C.A.I.-D., B.V., and M.A.N. supervised the study. J.G.R., A.P.M.-M., J.M.G., A.H., C.A.I.-D., B.V., and M.A.N. wrote the manuscript. All authors read and approved the manuscript. **Competing interests:** K.W.K. and B.V. are founders of Personal Genome Diagnostics. B.V. and K.W.K. are on the Scientific Advisory Board of Sysmex-Inostics. B.V. is also on the Scientific Advisory Boards of Exelixis GP. These companies and others have licensed technologies from Johns Hopkins, and K.W.K. and B.V. receive equity or royalties from these licenses. The terms of these arrangements are being managed by Johns Hopkins University in accordance with its conflict of interest policies. **Data and materials availability:** Accession nos. for the raw sequencing data are available in the original publications (*13*–*18*). Data of Brown *et al*. (*18*), Hong *et al*. (*14*), and Makohon-Moore *et al*. (*17*) as well as of subjects MSKA1 and MSKA2 are deposited at the European Genome-Phenome Archive (www.ebi.ac.uk/ega) and are available under accession nos. EGAS00001000760, EGAS00001000942, EGAS00001002186, and EGAS00001002777, respectively. Data of Gibson *et al*. (*16*) and Sanborn *et al*. (*13*) are deposited to the database of Genotypes and Phenotypes (dbGaP) at the National Center for Biotechnology Information (NCBI) under accession codes phs001127.v1.p1 and phs000941.v1.p1, respectively. Data of Kim *et al*. (*15*) are deposited to the Sequence Read Archive (SRA) at the NCBI under the project ID of PRJNA271316.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6406/1033/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S32
Table S1
References (*30*–*56*)

27 March 2018; accepted 2 August 2018
10.1126/science.aat7171

Make Your Research Hit the Target



Science Translational Medicine publishes peer-reviewed, cutting-edge biomedical research in the fields of cardiology, cancer, immunotherapy, infectious diseases and vaccines, bioengineering and devices, neurology and neurodegenerative diseases, obesity, diabetes and metabolic disorders, drug discovery, genomic medicine, imaging, stem cell therapy and regenerative medicine.

Submit your research today

Learn more at: **ScienceTranslationalMedicine.org**

Science Translational Medicine

AAAS

The Ho-Am Prize ignites the passion and innovative spirit for scientific discovery

We welcome your recommendations for the 2019 Ho-Am Prize candidates.

- Award Categories : Science, Engineering, Medicine
- Eligibility : Researchers of Korean heritage
- Submission Deadline : Oct. 31, 2018

* Learn more about the prize and nomination : **www.hoamprize.org**

To continue Samsung Founder Ho-Am Byung-Chull Lee's (1910-1987) efforts to maximize both human potential and public interest, Chairman Kun-Hee Lee of Samsung established the Prize in 1990. The Prize is awarded annually to outstanding researchers of Korean heritage around the world who have made important contributions to the advancement of science.

THE HO-AM PRIZE

HOAM
THE HO-AM FOUNDATION

SAMSUNG

NEW ENGLAND BioLabs® Inc.

be INSPIRED
drive DISCOVERY
stay GENUINE

Flexible Precision.

NEBNext Direct® Custom Ready Panels for NGS target enrichment

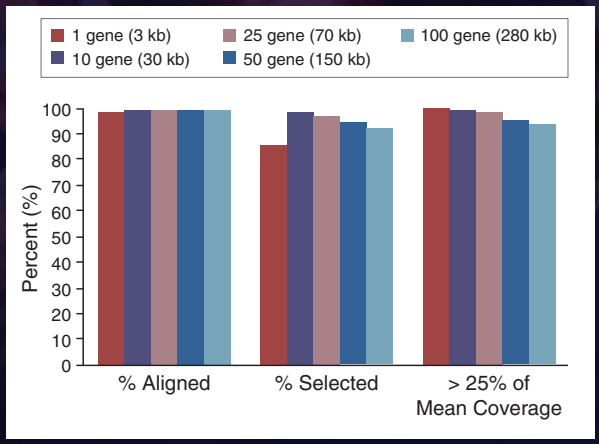
Employing the unique NEBNext Direct hybridization-based enrichment method, NEBNext Direct Custom Ready Panels allow rapid customization of targeted gene panels for Illumina® sequencing. Select from a list of genes for which baits have been carefully designed and optimized to give complete coverage of the full coding regions. High quality panels can be designed by you and rapidly delivered, from any combination of genes. NEBNext Direct Custom Ready Panels provide the content you want with the performance you need.

- Choose from a single gene to hundreds of genes
- Experience unmatched specificity and coverage uniformity
- Eliminate synthesis and optimization steps for faster turnaround
- Improve sensitivity with our Unique Molecule Index (UMI)
- Generate results in one day with our automation-friendly workflow

For more information visit

NEBNextDirect.com

NEBNext Direct Custom Ready Panels demonstrate optimum performance across a wide range of panel sizes



Key target enrichment metrics demonstrate consistent performance across a range of panel sizes. 100 ng of DNA was tested against panels of 1, 10, 25, 50 and 100 genes, and sequenced using Illumina® paired-end 150 bp sequencing. Larger panels included all genes present in smaller panels.

Present your work to the world

Are you a representative of the upcoming generation of thought leaders in your field? Together we look forward to your application for the new Sartorius & *Science* Prize for Regenerative Medicine & Cell Therapy.

Apply now!

www.passionforscience.com/prize

The Sartorius & *Science* Prize for Regenerative Medicine & Cell Therapy

Awarded by sartorius Science

new products: DNA/RNA analysis

Single Cell/ Low Input RNA Library Prep Kit

NEBNext is a unique streamlined workflow that allows the generation of the highest yields of high-quality, full-length transcript sequencing libraries from single cells, or from as little as 2 pg–200 ng total RNA. Enabling unmatched detection of low-abundance transcripts, this kit is compatible with a variety of RNA inputs, including cultured or primary cells or total RNA, and provides uniform and consistent transcript coverage, regardless of input amount or sample type.

New England Biolabs
For info: 800-632-5227
www.neb.com

Automated DNA–RNA Extraction

ExiPrep 16 Plus is a laboratory automation system able to extract DNA or RNA from various biological samples, including tissue, whole blood, serum, plasma, swab, bacteria, and yeast. It also features the ability to purify His-tagged recombinant proteins. ExiPrep 16 Plus can handle up to 16 samples at once, and includes optimized extraction protocols for many sample types so you get unsurpassed results from your nucleic acid or protein purification. It has advanced features, such as an intuitive touchscreen, a UV sterilization cycle, and a unique contamination shield, each designed to provide an extra level of confidence in your findings.

Bioneer
For info: 877-264-4300
us.bioneer.com

Blood DNA Extraction Kit

The smart Blood DNA Midi prep (a) is based on the unique SmartExtraction technology, and allows easy automation of DNA extraction from up to 3 mL of whole blood. Initially, erythrocytes are lysed and peripheral blood mononuclear cells are collected for further preparation. Prefilled, sealed reagent plastics simplify handling enormously and can be directly applied on InnuPure systems, GeneTheatre, or CyBio FeliX directly. Possible shearing effects on DNA are reduced to an absolute minimum due to SmartExtraction pipette tips, which contain modified surfaces for efficient binding of nucleic acids. Finally, high DNA yields (up to 90 µg from 3 mL whole blood) with high molecular weight (up to 500 kbp) can be isolated, with quality to ensure direct application in further downstream analyses.

Analytik Jena
For info: +49-(0)-3641-77-7444
www.analytik-jena.de/en

lncRNA RT-qPCR Workflow

Long noncoding RNAs (lncRNAs) do not code for proteins but make up most of the human transcriptome and are under active research as potential biomarkers and drug discovery targets. To study them efficiently, Bio-Rad has launched the lncRNA workflow, a real-time, quantitative PCR (RT-qPCR) workflow, which provides a streamlined, cost-effective alternative for lncRNA discovery and validation. The lncRNA workflow includes products that boost detection of low-abundance lncRNAs as well as experimentally validated primers to reliably detect and quantify more than 23,000 human lncRNA targets. For laboratories that find using RNA sequencing for lncRNA discovery expensive and time-consuming, the workflow provides 20 predesigned PrimePCR lncRNA arrays that cover the most frequently researched human lncRNAs.

Bio-Rad
For info: 510-741-1000
www.bio-rad.com/lncrnaworkflow

Nick Translation DNA Labeling System

The Nick Translation DNA Labeling System 2.0 provides a simple, efficient method for generating labeled DNA. The complete system includes all the necessary reagents required for 50 nick translation reactions and is recommended for labeling of double-stranded DNA larger than 1 kb for fluorescent in situ hybridization (FISH) applications. The system can accommodate a wide range of fluorophore-labeled, biotin-labeled, and digoxigenin-labeled nucleotides. In addition to choice of label, the kit design allows the user to optimize incorporation and product size by adjusting the ratio of labeled dUTP to dTTP. The ready-to-use NT Enzyme Mix is user friendly and minimizes pipetting errors. Probes labeled by nick translation can be used in many different hybridization techniques, including in situ hybridization, FISH, screening gene banks by colony or plaque hybridization, DNA or RNA transfer hybridization, and reassociation kinetics studies.

Enzo Life Sciences
For info: 800-942-0430
www.enzolifesciences.com

Next-Generation Library Preparation

Ion AmpliSeq HD is a library amplification technology for Ion Torrent next-generation sequencing that gives you the power to design your own gene panels and find variants with very low limits of detection—down to 0.1% for cell-free DNA samples. Use it to find low-frequency alleles in circulating tumor DNA, or within DNA or RNA samples extracted from formalin-fixed paraffin-embedded tissue, or to find rare microbial species or resistant strains. The heightened sensitivity achieved with Ion AmpliSeq HD is enabled with proprietary molecular tags, while dual barcoding of the fragments of interest helps enhance accuracy in fragment identification and sequencing using the Ion GeneStudio S5 Series instruments. The technology also provides workflow improvements that reduce library preparation time to less than 3 h.

Thermo Fisher Scientific
For info: 866-356-0354
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.

PCR

applied biosystems invitrogen

Nonstop results

Enhanced performance for first-class PCR

It's time to elevate your PCR

Together, Applied Biosystems™ and Invitrogen™ PCR products offer a direct route to reliable PCR results. Consistently arrive at results quickly, with more assurance and less optimization. So, sit back and relax, knowing we're here to help you get to your final PCR destination.

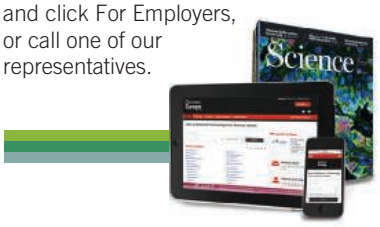
Upgrade your PCR at **thermofisher.com/amplifly**

ThermoFisher SCIENTIFIC

Science Careers

SCIENCE CAREERS ADVERTISING

For full advertising details, go to ScienceCareers.org and click For Employers, or call one of our representatives.



AMERICAS

+1 202 326-6577
+1 202 326-6578
advertise@sciencecareers.org

EUROPE, INDIA, AUSTRALIA, NEW ZEALAND, REST OF WORLD

+44 (0) 1223 326527
advertise@sciencecareers.org

CHINA, KOREA, SINGAPORE, TAIWAN, THAILAND

+86 131 4114 0012
advertise@sciencecareers.org

JAPAN

+81 3-6459-4174
advertise@sciencecareers.org

CUSTOMER SERVICE

AMERICAS

+1 202 326-6577

REST OF WORLD

+44 (0) 1223 326528

advertise@sciencecareers.org

All ads submitted for publication must comply with applicable U.S. and non-U.S. laws. *Science* reserves the right to refuse any advertisement at its sole discretion for any reason, including without limitation for offensive language or inappropriate content, and all advertising is subject to publisher approval. *Science* encourages our readers to alert us to any ads that they feel may be discriminatory or offensive.

ScienceCareers
FROM THE JOURNAL SCIENCE AAAS

ScienceCareers.org

Career Feature:

Artificial Intelligence

Issue date: November 30

Book ad by November 15

Ads accepted until November 21 if space allows

129,562
subscribers in print every week

503,472
monthly unique browsers on ScienceCareers.org

56 %
of our weekly readers are Ph.D.s

Artificial Intelligence (AI) is impacting science in new and exciting ways as scientists are using it to better understand society to find solutions to problems across diverse disciplines. This feature will give an overview of AI, and explore the hotspots/centers of excellence and applications for AI. Typical career paths for those working in AI will be explored as well as the opportunities that exist for careers in AI.

Your organization can brand itself as a leader in AI by raising your visibility alongside relevant content while attracting potential candidates. Contact us for further details.

To book your ad:
advertise@sciencecareers.org

The Americas
+ 202 326 6577

Europe
+44 (0) 1223 326527

Japan
+81 3 6459 4174

China/Korea/Singapore/ Taiwan
+86 131 4114 0012

Produced by the *Science*/AAAS Custom Publishing Office.

What makes *Science* the best choice for recruiting?

- Read and respected by 400,000 readers around the globe
- Your ad dollars support AAAS and its programs, which strengthens the global scientific community.

Why choose this AI Feature for your advertisement?

- Relevant ads lead off the career section with a special "AI" banner.

Expand your exposure by posting your print ad online:

- Link on the job board homepage directly to AI jobs
- Dedicated landing page for AI positions.

ScienceCareers
FROM THE JOURNAL SCIENCE AAAS

SCIENCECAREERS.ORG

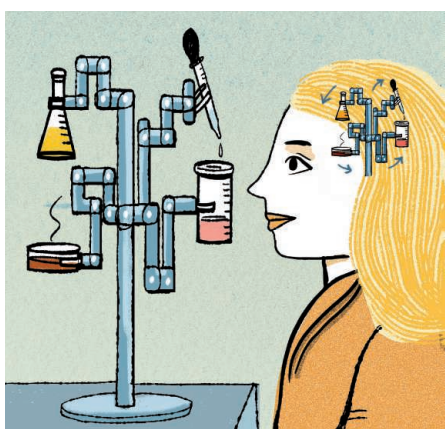
FOR RECRUITMENT IN SCIENCE, THERE'S ONLY ONE *SCIENCE*.

By **Sarah Anderson**

Outsmarting our instruments

I'm a graduate student in a lab that seemingly has an instrument for everything. Peptide synthesizer? We've got it. Liquid-handling instrument for 96-well plates? No problem. You name it, we've got a robot that can do it. The convenience and efficiency can't be beat. But when I first joined the lab, I feared that these tools would make grad students like me obsolete. I thought the single quality that defined a great scientist was perfect experimental technique, and that scientists are essentially supposed to function as living, breathing instruments. Gradually, though, I accepted that these instruments had better experimental technique than I ever could master, and that I had better rethink what it really means to be a scientist.

My misconceptions, and my anxiety around them, arose during my undergraduate training. I excelled at homework assignments and exams, but I often struggled in the lab. In my freshman-year chemistry lab, for example, we were supposed to determine the concentration of a hydrochloric acid solution by meticulously adding sodium hydroxide until the clear liquid turned pink—a standard acid-base titration. When done correctly, it's beautiful. At first, you're just combining two colorless solutions. Then, each drop of sodium hydroxide creates an ethereal swirl of pink that gradually dissipates as the solution becomes clear again. Finally, with just one more drop of sodium hydroxide, the solution blinks from clear to all-over pink.



"I feared that these tools would make grad students like me obsolete."

But my hands weren't steady enough to precisely dispense the sodium hydroxide. I shot right past the endpoint, turning the contents of my beaker an embarrassingly vivid magenta. As I looked around the lab at the sea of perfectly pale pink beakers achieved by students with better reflexes, I began to doubt whether I was cut out to be a scientist. No one, I thought, would care that I could predict the product of a reaction if I couldn't successfully run it in the lab.

Even so, I continued with my chemistry major, driven by my passion for the material and positive feedback about my performance in the lecture components of my classes. I was also fortunate to have a professor who told me that the best predictor of success was whether I could master the concepts—which I could—and that, with practice, my experimental technique would not be an insurmountable barrier. She also insisted that grad school was the right option for me. I wasn't entirely convinced. But I had hoped to pursue a Ph.D. ever since I started my undergrad degree, so I decided to give it a try.

There, I slowly came to understand what my mentor meant: Being a scientist is much more about creativity than about what we can do with our hands. This really started to sink in during my second year, when I adapted an experimental protocol for use in living cells. I needed to design a molecule that would be able to cross a cell membrane and then come up with a way to synthesize it. There's no instrument for that. I dusted off my old medicinal chemistry notebook, compiled a list of potential structures, searched the literature, and homed in on my target.

I did eventually have to do the experiments, including synthesizing the molecule—which required several attempts, many helpful tips from a senior graduate student, and countless absorbent spill pads. But ultimately, I got the off-white powder I needed, and I managed to improve my experimental technique along the way. Only then did I turn to the state-of-the-art instruments at my disposal, which I had come to appreciate as tools to support my work rather than things to strive to emulate.

Now, whenever I am faced with a challenging experimental task and feel that freshman-year doubt in my scientific prowess begin to creep in, I think of Alexander Fleming. A technically great scientist would have thrown out a moldy bacteria culture to prevent contamination. But as a truly great scientist, he took the time to consider the phenomenon—and used it to invent penicillin. And I bet Sir Alexander wasn't great at titrations either. ■

Sarah Anderson is a Ph.D. candidate at Northwestern University in Evanston, Illinois. Send your career story to SciCareerEditor@aaas.org.