

A user's guide to the new  
world of open access p. 16

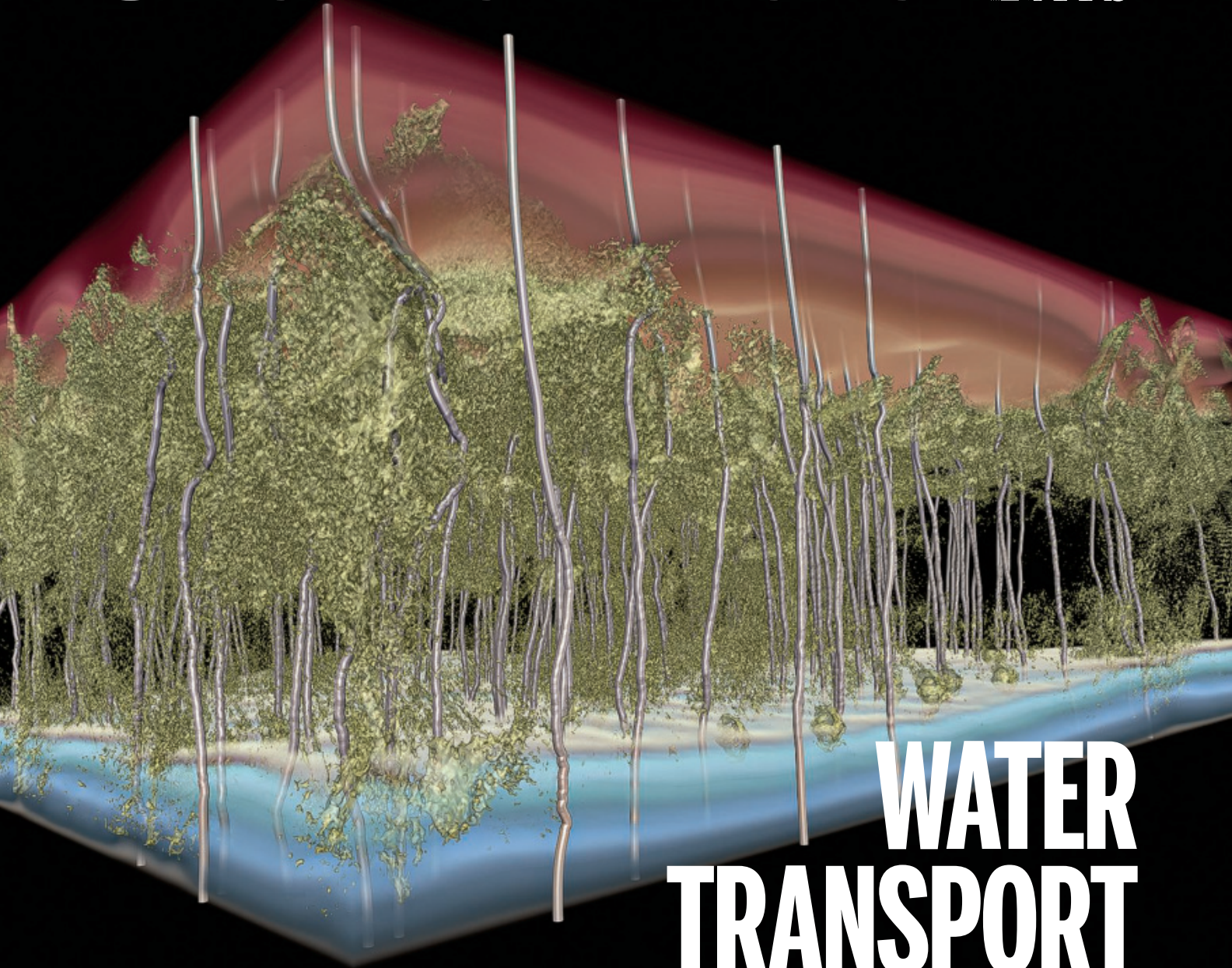
Young scientists coin words  
to describe 2020 p. 22

Probing the complexities of  
clock-cancer interactions p. 42

# Science

\$15  
1 JANUARY 2021  
[sciencemag.org](http://sciencemag.org)

AAAS



## WATER TRANSPORT

Membrane homogeneity maximizes flow  
without sacrificing desalination pp. 31 & 72



# TRILLIONS OF MICROBES ONE ESSAY

The NOSTER *Science* Microbiome Prize is an international prize that rewards innovative research by investigators, under the age of 35, who are working on the functional attributes of the microbiota. The research can include any organism that has potential to contribute to our understanding of human or veterinary health and disease, or to guide therapeutic interventions. The winner and finalists will be chosen by a committee of independent scientists, chaired by a senior editor at *Science*. The top prize includes a complimentary membership to AAAS, an online subscription to *Science*, and \$25,000 (USD). Submit your research essay today.



Oliver Harrison  
2020 Grand Prize Winner

**NOSTER** | *Science*  
**MICROBIOME**  
PRIZE

Apply by 1/24/21 at [www.sciencemag.org/noster](http://www.sciencemag.org/noster)

Sponsored by Noster Inc.



# CONTENTS

1 JANUARY 2021 • VOLUME 371 • ISSUE 6524



16

## NEWS

### IN BRIEF

#### 6 Areas to watch

What's coming up in 2021

EDITORIAL p. 5; PODCAST

### IN DEPTH

#### 9 Fast-spreading U.K. virus variant raises alarms

Scientists are scrambling to better understand effects of a series of worrisome mutations *By K. Kupferschmidt*

#### 10 Pfizer's vaccine raises allergy concerns

Polymer in mRNA's "packaging" may cause rare anaphylactic reactions *By J. de Vrieze*

#### 12 A health economist confronts Kenya's pandemic

Edwine Barasa helps guide the government's response with data—and quiet persistence *By L. Nordling*

#### 13 Alaska oil bid alarms scientists

Mapping plan for Arctic refuge ignores risks, critics say *By W. Cornwall*

#### 14 Slowdown in plate tectonics may have led to ice sheets

Decreased ocean crust production tied to rapid cooling *By P. Voosen*

#### 15 Congress backs research in 2021 spending bill

Modest increases for science complete 4 years of substantial growth despite Trump *By J. Mervis*

### FEATURES

#### 16 Open access takes flight

As a new mandate takes effect, researchers and institutions grapple with the trade-offs of making scientific publications free for all *By J. Brainard*



22

## INSIGHTS

### LETTERS

#### 22 NextGen Voices

Defining events: 2020 in hindsight

### PERSPECTIVES

#### 25 COVID-19 and cancer in Africa

The impacts of COVID-19 present substantial challenges and opportunities in global oncology  
*By B. W. Addai and W. Ngwa*

#### 27 The puzzle of the COVID-19 pandemic in Africa

More data are needed to understand the determinants of the COVID-19 pandemic across Africa  
*By J. M. Maeda and J. N. Nkengasong*  
REPORT p. 79

#### 29 RNA-targeted drugs for neuromuscular diseases

Progress with antisense oligonucleotide therapies opens a path for future development *By A. Ferlini et al.*

#### 31 Why polyamide reverse-osmosis membranes work so well

Inhomogeneities in membrane thickness and density promote water transport  
*By G. M. Geise*  
REPORT p. 72

#### 32 Detecting oxygen changes in the lungs

Lung airway basal stem cells directly sense changes in oxygenation, driving lung regeneration  
*By W. Zacharias*  
RESEARCH ARTICLE p. 52

### POLICY FORUM

#### 34 Mapping the global threat of land subsidence

Nineteen percent of the global population may face a high probability of subsidence  
*By G. Herrera-García et al.*

### BOOKS ET AL.

#### 37 Science's irrational origins

Disputes in modern science are settled with empiricism alone, an approach early scholars would have questioned  
*By I. Yanai and M. J. Lercher*

#### 38 Thermodynamics and the matter of life

A scientist considers life's genesis through the physics of Exodus  
*By E. A. Mukamel and A. M. Glaser*



37

## RESEARCH

### IN BRIEF

**39** From *Science* and other journals

### REVIEW

#### **42** Circadian rhythms

Clocks, cancer, and chronochemotherapy

A. Sancar and R. N. Van Gelder

REVIEW SUMMARY; FOR FULL TEXT:  
DOI.ORG/10.1126/SCIENCE.ABB0738

### RESEARCH ARTICLES

#### **43** Structural biology

Structural basis of antagonizing the vitamin K catalytic cycle for anticoagulation

S. Liu et al.  
RESEARCH ARTICLE SUMMARY; FOR FULL TEXT:  
DOI.ORG/10.1126/SCIENCE.ABC5667

#### **44** Transcription

Steps toward translocation-independent RNA polymerase inactivation by terminator ATPase  $\rho$

N. Said et al.  
RESEARCH ARTICLE SUMMARY; FOR FULL TEXT:  
DOI.ORG/10.1126/SCIENCE.ABD1673

#### **45** Stress responses

QRICH1 dictates the outcome of ER stress through transcriptional control of proteostasis

K. You et al.  
RESEARCH ARTICLE SUMMARY; FOR FULL TEXT:  
DOI.ORG/10.1126/SCIENCE.ABB6896

#### **46** Batteries

A rechargeable zinc-air battery based on zinc peroxide chemistry

W. Sun et al.

#### **52** Stem cells

Airway stem cells sense hypoxia and differentiate into protective solitary neuroendocrine cells

M. Shivaraju et al.  
PERSPECTIVE p. 32

#### **57** Protein synthesis

Interactions between nascent proteins translated by adjacent ribosomes drive homomer assembly

M. Bertolini et al.

### REPORTS

#### **Cell cycle**

**64** A tripartite mechanism catalyzes Mad2-Cdc20 assembly at unattached kinetochores

P. Lara-Gonzalez et al.  
**67** CDC20 assists its catalytic incorporation in the mitotic checkpoint complex

V. Piano et al.

#### **72** Membranes

Nanoscale control of internal inhomogeneity enhances water transport in desalination membranes

T. E. Culp et al.  
PERSPECTIVE p. 31

#### **76** Materials science

Achieving large uniform tensile elasticity in microfabricated diamond

C. Dang et al.

#### **79** Coronavirus

Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Kenyan blood donors

S. Uyoga et al.  
PERSPECTIVE p. 27

#### **83** Cloud physics

Aerosol invigoration of atmospheric convection through increases in humidity

T. H. Abbott and T. W. Cronin

#### **86** Protein folding

Evolution of fold switching in a metamorphic protein

A. F. Dishman et al.

#### **90** Active matter

Low rattling: A predictive principle for self-organization in active collectives

P. Chvykov et al.

### DEPARTMENTS

#### **5** Editorial

A little better all the time in 2021

By H. Holden Thorp

NEWS p. 6

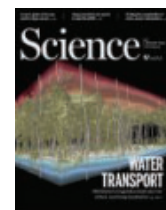
#### **98** Working Life

Saying yes to help

By Angela Q. Zhang

### ON THE COVER

Multimodal electron microscopy reveals the three-dimensional nanostructure (gold) of reverse-osmosis membranes, as well as the need to control their polymer mass distribution for improved performance. Volume reconstructions with nanometer resolution are used as inputs for water flow simulations that reveal streamlines (gray; here, water flows from top to bottom). Minimizing “dead zones,” thereby achieving a more uniform density throughout the membranes, is key to maximizing water production. See pages 31 and 72. *Data Visualization:* Greg Foss, Texas Advanced Computing Center, University of Texas, Austin



Science Careers .....96

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. Copyright © 2021 by the American Association for the Advancement of Science. The title SCIENCE is a registered trademark of the AAAS. Domestic individual membership, including subscription (12 months): \$165 (\$74 allocated to subscription). Domestic institutional subscription (51 issues): \$2148; Foreign postage extra: Air assist delivery: \$98. 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 available from backissues.science.org; 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 can be obtained through the Copyright Clearance Center (CCC), www.copyright.com. The identification code for Science is 0036-8075. Science is indexed in the Reader's Guide to Periodical Literature and in several specialized indexes.



# A little better all the time in 2021

A famous story about the Beatles tells of the collaboration between Paul McCartney and John Lennon on the song “Getting Better” for their legendary Sgt. Pepper’s Lonely Hearts Club Band album. After McCartney wrote the lines “I’ve got to admit, it’s getting better; a little better all the time,” Lennon wryly added, “It can’t get no worse.” This story could serve as an epigraph as the calendar turns from the year 2020, which could hardly have gotten much worse, to 2021, when we hope life will indeed get a little better all the time. Better from COVID-19 because of the vaccines, better from misinformation spread by outgoing president Donald Trump and his allies, and better, we can hope, when it comes to the production and distribution of scientific knowledge.

There’s plenty of exciting science to be optimistic about in 2021 (see News on p. 6). At the end of 2020, the DeepMind group in the United Kingdom announced a major advance in long-standing challenges in protein folding, predicting three-dimensional (3D) structures of proteins from their amino acid sequence. The next year portends even more exciting advances in protein structure and design.

On the cosmic front, there are many efforts underway to bring samples from the Solar System back to this planet. The Hayabusa2 mission that traveled to the asteroid 162173 Ryugu retrieved what could be a treasure trove of material revealing details about the ancient delivery of water and organic molecules to Earth. Similarly, the OSIRIS-REx (Origins, Spectral Interpretation, Resource Identification, Security, Regolith Explorer) mission has collected samples from the asteroid Bennu that, when they arrive, could reveal important aspects of the formation of the Solar System. The new Mars rover Perseverance will land in February and, in addition to transmitting important data from the red planet, will begin the process of collecting samples that may eventually be studied in terrestrial laboratories.

In biology, the COVID-19 pandemic led to major advances in the development and application of messenger RNA (mRNA) vaccines. It is stunning that science not only came up with a vaccine to a new pathogen so quickly but also advanced a brand new vaccine tech-

nology, albeit one that was already in development for several years. The application of mRNA therapies to other problems in infectious diseases and throughout medicine will be exciting to follow.

Quantum computing remains an important area to watch. This year, *Science* published a paper that describes the application of a quantum computer to an important problem in theoretical chemistry. In the coming months, it’s likely that there will be progress in addressing the problem of quantum error correction, pushing quantum computing a little closer to routine application.

Additive manufacturing and 3D printing continue to become more practical. In particular, the ability to apply these techniques to new types of materials will

make it more likely that advanced manufacturing can benefit from the science behind these processes.

On the policy front, the continued development of the UK Research and Innovation (UKRI) organization—as described in a recent editorial by Ottoline Leyser—will be of keen interest as the Brexit process continues. Despite choppy politics, the scientific vision of UKRI is strong and could lead to advances in British science.

In the United States, although the Biden White House will certainly be friendlier to science, the science denial that fueled the Trump administration will linger in the American population and among some conservative politi-

cians. The battles ahead are not to be underestimated. Continued denial of climate change and COVID-19 is sadly inevitable, and it will take everything U.S. science and the Biden administration can muster to stay strong. Still, as new leaders are named and confirmed in health and science policy, U.S. science should be able to at least catch its breath and feel optimistic about a new era.

Although 2020 will certainly go down as a year that couldn’t get much worse, there is plenty to be proud of and reason to hope that things will be getting better. The virus was confronted. Epidemiologists and other scientists became household names. And the scientific community found a much stronger voice, one that will serve us all well in 2021 and beyond.

– H. Holden Thorp



**H. Holden Thorp**  
is Editor-in-Chief,  
*Science Journals*.  
hthorp@aaas.org;  
@hholdenthorp

“...U.S. science  
should be  
able to at least  
catch its  
breath and feel  
optimistic about  
a new era.”



# NEWS

## IN BRIEF

Edited by Jeffrey Brainard

### AREAS TO WATCH

## What's coming up in 2021

**A**s biomedical scientists continue to battle the deadly pandemic this year to help the world return to normalcy, researchers across the disciplines still aim to hit big milestones or launch new projects despite the challenges brought by COVID-19. European scientists will also have to contend with the aftermath of Brexit. Many U.S. scientists, in contrast, have a more hopeful political outlook, with some likely to play an invigorated role in tackling another global crisis, climate change, after President-elect Joe Biden, who has vowed to make it a top priority, is sworn in this month. In this section, *Science's* news staff forecasts areas of research and policy we expect to make headlines this year, from protecting the high seas' biodiversity to probing how ancient humans interacted.

### Climate change sized up

**GLOBAL WARMING** | Nearly 8 years have passed since the fifth assessment report from the United Nations's Intergovernmental Panel on Climate Change, the famed body of volunteer climate scientists that since 1990 has chronicled humanity's persistent warming of the planet. The sixth installment, crafted by more than 700 scientists and delayed by the pandemic, will come out in sections this year and next, and it's expected to further sharpen the picture of the human impact on climate. Its findings

will be bolstered by a new generation of climate models and scenarios, fed by indicators of unabated global change: data showing accelerating sea level rise; rapid ice melt at the poles; and waves of extreme heat, drought, and fire. In November, the world's countries will meet in Scotland's Glasgow for the next U.N. climate summit, where they are expected to increase the ambitions of their pledged cuts in greenhouse gases and agree on a full set of rules to implement the Paris agreement. Among the attendees will be the United States, which President-elect Joe Biden has said will rejoin the pact.

### WHO probes coronavirus origin

**PUBLIC HEALTH** | The launch of immunization campaigns in many countries has raised hopes that the COVID-19 pandemic can be brought to an end. But how exactly it started remains murky, and a World Health Organization international team of 10 scientists will travel to China several times this year as part of an investigation into the pandemic coronavirus' origins—a politically sensitive mission because the United States and China have sparred over who is to blame for the pandemic. The team hopes to discover the virus' closest relatives in bats, where and how it jumped to humans, whether another species acted as an intermediate host, and, most important, how we can prevent other pandemic viruses from emerging.

### Repairing frayed U.S.-Chinese ties

**NATIONAL SECURITY** | Restoring U.S.-Chinese scientific collaborations to health could be a test of President-elect Joe Biden's success in negotiating with the Asian superpower on trade, immigration, and security concerns. A new U.S. government-sponsored forum on science, technology, and national security will advise the new administration on how to strike the right balance between openness and preventing the theft of new technology. A new U.S. law provides stiff penalties for federal grant applicants who fail to disclose





Climate change is expected to bring more droughts that worsen blazes like California's Silverado Fire in 2020.

all sources of funding when submitting a proposal, based on the assumption that greater transparency is the best way to monitor ties with China and other entities posing potential threats to U.S. security. University leaders are hoping Biden will reverse President Donald Trump's restrictive immigration policies and avoid ratcheting up economic and political tensions. But prominent Republicans in Congress are threatening even more punitive steps against China.

## Drugs tailored for COVID-19

**INFECTIOUS DISEASES** | To complement mass vaccinations against COVID-19, drug companies this year will dash to custom design drugs that block the pandemic coronavirus and treat the disease's symptoms. Even if many people receive one of the new, highly effective vaccines approved in recent weeks by regulators, the virus is expected to remain endemic. In 2020, only the antiviral remdesivir and a handful of other drugs, all originally designed to treat other conditions, showed even limited benefits against the disease. To identify new drug candidates, researchers have deployed artificial intelligence and supercomputers, and more than 590 experimental drugs are in development, according to a leading pharmaceutical industry tracker. For example, researchers have high hopes for compounds that disrupt reproduction of the pandemic virus by inhibiting one of its

two proteases. Cocktails of such therapies could tame the virus, an approach used successfully against HIV. The protease inhibitors and other compounds look promising in cell and animal studies. But studies on human volunteers are only getting started, and it could take years to pass safety and efficacy reviews.

## New rovers for Red Planet

**PLANETARY SCIENCE** | Mars's thin air makes it hard to slow a probe to a soft landing. Of the 18 robotic probes sent to the planet's surface in the past 50 years, eight have crashed. This year, two more will attempt a touchdown. On 18 February, NASA's SUV-size rover, Perseverance, will take the plunge, slowed by parachutes and retrorockets on a "sky crane" platform. After landing in Jezero crater, near a fossilized river delta, the rover will collect rock samples for eventual return to Earth. Around the same time, China's Tianwen-1 mission will arrive with an orbiter, a lander platform, and a rover the size of a golf cart. Officials have chosen a landing site not far from Jezero, along the southern edge of Utopia Planitia, a broad plain that may have been repaved by ancient flows of mud. A successful touchdown would be China's first on Mars.



A "sky crane" landing device is designed to lower a 1-ton NASA rover to a soft landing on Mars.

## A sharper view of proteins

**MICROSCOPY** | Researchers aim this year to sharpen the resolution of cryo-electron microscopy (cryo-EM), a technique for studying protein structure that may yield new insights into their roles in maintaining human health and causing disease. Another technique, x-ray crystallography, has long been the gold standard for mapping individual atoms within a 3D protein structure. But it only works for proteins that can be packed into crystals. Cryo-EM doesn't require crystals, and its resolution has steadily improved over the past decade. In 2020, it crossed the threshold of atomic resolution as researchers using cryo-EM microscopes equipped with improved electron detectors and software mapped the structure of apoferritin, an iron-binding protein. That protein is unusually rigid, which made it easier to hold steady during cryo-EM mapping. Next, researchers want to image less rigid proteins. Success would be a boon to structural biologists, allowing them to generate highly detailed maps of large proteins and complexes of multiple proteins that cannot be crystallized.

## Webb telescope nears launch

**ASTRONOMY** | The long wait will soon be over: NASA's much-delayed flagship observatory, the James Webb Space Telescope (JWST), is set to finally take to the skies on 31 October. JWST is the successor to the Hubble Space Telescope, with a 6.5-meter-wide mirror that has six times its predecessor's light-gathering power. The gold-coated, honeycombed mirror will be cooled so it can collect the infrared light of distant objects, red-shifted by the universe's expansion. JWST will be sensitive enough to scrutinize the atmospheres of nearby exoplanets for signs of life and gather the light of the universe's first stars and galaxies. The \$8.8 billion spacecraft, which will cost billions more and launch years later than originally planned, recently endured final tests—violent shaking to simulate launch. This month, engineers are unfolding its mirror and unfurling its multilayered Sun shield one last time to check that all is well. By midyear, JWST will be packed up and shipped to French Guiana, where it will be loaded onto a European Ariane 5 rocket. Next stop: deep space.

## Reactor aims for energy gain

**ENERGY** | The Joint European Torus (JET), the world's largest fusion reactor, will this year embark on a campaign to generate



substantial amounts of fusion power. The U.K.-based JET is a tokamak, which uses powerful magnets to constrain a hot plasma so that atomic nuclei crash together and fuse, releasing energy. After an upgrade, JET has a new metallic lining and extra heating power; in this year's trials, it will be fed a potent mix of the hydrogen isotopes deuterium and tritium (D-T)—a fuel rarely used because radioactive tritium needs careful handling and cleanup. In 1997, the last time this fuel mix was used, JET generated 16 megawatts of power for a few seconds, well short of the power consumed to make it happen. The new campaign will initially aim for similar power levels but try to sustain them for longer. That will help in planning for the huge ITER reactor, under construction in France, which has a similar shape and lining. ITER is due to start operations in 2025, but will not begin to use D-T fuel until the mid 2030s.

## Gut health for malnourished kids

**NUTRITION** | Help may arrive this year for millions of malnourished children who remain sickly and fail to fully recover even after receiving proper nutrition and treatments for being underfed. Pandemic-related disruptions and job losses are expected to cause the number of such children to skyrocket. One problem for underfed children is disruption of the gut microbiome, which leads to an immature, inefficient digestive system that stunts growth. To repair the microbiomes, health specialists are looking forward to results of a study in Bangladesh that evaluated a low-cost nutritional supplement; it's a mix of easy-to-find ingredients, such as chickpeas, bananas, and soy and peanut

flours. In 2019, this team reported that in experiments in mice and pigs, followed by a monthlong pilot study of 60 children, the supplement led to gut repair, as indicated by changes in blood markers. The study did not last long enough to test effects on growth. Since then, these researchers have been comparing the intervention with an existing supplement in a larger, 3-month trial of malnourished children.

## U.N. to protect high seas

**CONSERVATION** | Few rules protect biodiversity in the two-thirds of the ocean that lies outside nations' sovereign waters. This year, the United Nations is expected to finalize the first treaty specifically intended to change that. The pact is expected to provide a way to designate marine protected areas (MPAs) on the high seas; researchers have been developing a list of candidates and supporting evidence. The draft language also sets minimum standards for environmental impact assessments that nations would be required to conduct before starting commercial activities that might harm marine life. A new, international scientific and technical body, similar to one that manages marine life around Antarctica, would review MPA proposals. The treaty draft also provides for a system to manage genetic sequences taken from marine organisms living in the high seas.

## New clues to ancient societies

**ARCHAEOLOGY** | Expect to see studies of ancient humans follow new paths this year, as researchers combine analyses of ancient DNA with other molecular and

microbial clues to examine social ties and migrations. Scientists will merge DNA evidence with data from proteins and isotopes, as well as microfossils and pathogens from bones, tooth plaque, and fossilized poop. Such studies this year could help determine which early Celtic family members inherited wealth. They could help identify the homeland of the biblical Philistines and clarify the identities of early Anglo-Saxons and Greeks in Europe, as well as mummies in China and Egypt.

## Stingier vaccine injury payments

**PUBLIC HEALTH** | An obscure \$4 billion U.S. government fund that compensates people injured by vaccines is poised to become more tight-fisted this month. Changes proposed by the Trump administration will likely take effect in mid-January, making it more complicated and time consuming for people to win a payout if they sustain shoulder injuries after incorrectly administered injections with flu, tetanus, and other vaccines. The new rules won't affect people who might be injured by COVID-19 vaccines, who would need to apply to a different government program for compensation. But claims that other vaccines caused shoulder injuries have grown on the heels of expanded influenza vaccination and a 2010 paper in which government scientists first described "shoulder injury related to vaccine administration."

## Cancer drug nears approval

**BIOMEDICINE** | For more than 3 decades, scientists have dreamed of shrinking tumors by shutting off a protein called KRAS whose growth signals drive many cancer types. KRAS was thought to be impervious to drugs, in part because it offered no obvious pockets that inhibitors could target. But multiple companies have now developed compounds that fit into a groove on some cancer-promoting mutant KRAS proteins and curb their signals. The drugs have shown promising results, first in rodents and then in cancer patients. In December 2020, Amgen asked the Food and Drug Administration to review its KRAS drug, sotorasib, setting the stage for approval this year of the first member of this novel drug class. The drug could first be licensed for use in certain lung cancer patients. Another firm is expected to submit its KRAS drug for approval this year as well.



A new treatment could help children suffering from malnutrition, such as this boy at a treatment center for Rohingya refugees in Bangladesh.

**S** [SCIENCEMAG.ORG/NEWS](https://www.sciencemag.org/news)  
Read more news from Science online.



Trucks on their way to France queue in southeastern England on 21 December 2020 after the border was closed in an attempt to stop the spread of a new SARS-CoV-2 variant.

## COVID-19

# Fast-spreading U.K. virus variant raises alarms

Scientists are scrambling to better understand effects of a series of worrisome mutations

By Kai Kupferschmidt

**O**n 8 December 2020, a small group of scientists in the United Kingdom logged on for a regular Tuesday videoconference about the spread of the pandemic coronavirus. The discussion focused on Kent, a county in southeastern England that was seeing increasing transmission of SARS-CoV-2, even as the rest of the country was managing to curb the spread. Because investigations had not found any obvious causes—no big workplace outbreaks or changes in people's behavior—several researchers had been asked to look at viral genomes from the region.

The genetic family tree they presented showed something unusual was going on, says one of the attendees, microbial genomicist Nick Loman of the University of Birmingham. Not only were half the cases in Kent caused by a specific variant of SARS-CoV-2, but its branch literally stuck out from the rest of the data. "I've not seen a part of the tree that looks like this before," Loman says. And when scientists compared how fast this variant, named B.1.1.7, and others were spreading, they made a startling discovery: The virus seemed to have become more adept at transmitting between people.

The discovery of the viral lineage, along with a similarly worrying one in South Af-

rica, had a massive impact. On 19 December, U.K. Prime Minister Boris Johnson announced that London and southeastern England would be placed under tighter COVID-19 restrictions to contain the variant, which Johnson said may be 70% more transmissible. Although there's no evidence yet that the strain is more deadly, many countries closed their borders to travelers from the United Kingdom as they mulled how to deal with the possible new threat. Several announced they, too, had the variant among their populations.

As this issue of *Science* went to press on 23 December, scientists were still grappling to understand whether the variant really spreads faster, and if so, how. But its emergence had driven home the notion that viral evolution, which so far has had little impact on the trajectory of the COVID-19 pandemic, could yet result in nasty surprises—just as the first effective vaccines are being rolled out. It also raises the question of whether those vaccines may need periodic updating to parry a changing virus.

The U.K. lineage of SARS-CoV-2 has apparently acquired 17 mutations that lead to amino acid changes in its proteins all at once, a feat never seen before in the coronavirus. Crucially, eight of them were

in the gene that encodes spike, a protein on the viral surface that the pathogen uses to enter human cells. "There's now a frantic push to try and characterize some of these mutations in the lab," says Andrew Rambaut, a molecular evolutionary biologist at the University of Edinburgh.

Three already stand out as worrisome. A mutation named N501Y has previously been shown to increase how tightly spike binds to the angiotensin-converting enzyme 2 receptor, its main entry point into human cells. Scientists in South Africa were the first to spot N501Y's importance: They noted it several weeks ago in a lineage that is surging in the Eastern Cape, Western Cape, and KwaZulu-Natal provinces. "We found that this lineage seems to be spreading much faster,"

says Tulio de Oliveira, a virologist at the University of KwaZulu-Natal whose work alerted U.K. scientists to the mutation. That's concerning, says evolutionary biologist Jesse Bloom of the Fred Hutchinson Cancer Research Center: "Anytime you see the same mutation being independently selected multiple times, it increases the weight of evidence that that mutation is probably beneficial in some way for the virus."

B.1.1.7's second notable mutation, a deletion named 69-70del, leads to the loss of two

*Science's*  
COVID-19  
reporting is  
supported by the  
Pulitzer Center  
and the  
Heising-Simons  
Foundation.



amino acids in the spike protein. It, too, had appeared before: It was found, together with another mutation named D796H, in the virus of a COVID-19 patient in Cambridge, U.K., who was given plasma from recovered patients as a treatment, but eventually died. In lab studies, the patient's strain was less susceptible to convalescent plasma from several donors than wild-type virus, says Ravindra Gupta, a virologist at the University of Cambridge who published the findings in a preprint in early December.

Gupta also engineered a lentivirus to express mutated versions of SARS-CoV-2's spike and found that the deletion alone made the virus twice as infectious for human cells. A third mutation, P681H, is one to watch as well, says virologist Christian Drosten of the Charité University Hospital in Berlin, because it changes the site where the spike protein is cleaved before it enters human cells.

New virus strains are common in outbreaks and often spark alarm, but few are ultimately consequential. So U.K. scientists and others were initially cautious about concluding that B.1.1.7's mutations made the virus better at spreading from person to person. But the new variant is rapidly replacing other viruses, says Müge Çevik, an infectious disease specialist at the University of St. Andrews. Yet exactly what impact each mutation has is much more difficult to assess than spotting them or showing they're on the rise, says Seema Lakdawala, a biologist at the University of Pittsburgh.

Animal experiments can help show an effect, but they have limitations. Hamsters already transmit SARS-CoV-2 virus rapidly, for instance, which could obscure any effect of the new variant. Ferrets transmit it less efficiently, so a difference may be more easily detectable, Lakdawala says. "But does that really translate to humans? I doubt it." A definitive answer may be months off, she predicts.

The slew of mutations also raised worries that the South African or U.K. lineage might lead to more severe disease or even evade vaccine-induced immunity. So far there is little reason to think so. Whereas some mutations have been shown to let the virus evade monoclonal antibodies, vaccines and natural infections both appear to lead to a broad immune response that targets many parts of the virus, says Shane Crotty of the La Jolla Institute for Immunology. "It would be a real challenge for a virus to escape from that." The

measles and polio viruses have never learned to escape the vaccines targeting them, he notes: "Those are historical examples suggesting not to freak out."

At a 22 December press conference, BioNTech CEO Uğur Şahin pointed out that the U.K. variant differed in only nine of more than 1270 amino acids of the spike protein encoded by the messenger RNA in the very effective COVID-19 vaccine his company developed with Pfizer. "Scientifically it is highly likely that the immune response by this vaccine also can deal with the new virus," he said. Experiments are underway that should soon confirm that, Şahin added.



The mutation N501Y affects amino acids (yellow) in the spike protein, which binds to a human receptor (green).

Another major question is how the virus accumulated a host of mutations in one go. So far, SARS-CoV-2 typically acquired only one to two mutations per month. Scientists believe the new variant may have gone through a lengthy bout of rapid evolution in a chronically infected patient who then transmitted the virus. "We know this is rare but it can happen," says World Health Organization epidemiologist Maria Van Kerkhove.

Sébastien Calvignac-Spencer, an evolutionary virologist at the Robert Koch Institute, says the United Kingdom's new COVID-19 lockdown and other countries' border closures mark the first time such drastic action has been taken based on genomic surveillance in combination with epidemiological data. "It's pretty unprecedented at this scale," he says. But the question of how to react to disconcerting mutations in pathogens will crop up more often, he predicts. Most people are happy they prepared for a category 4 hurricane even if the predictions turns out to be wrong, Calvignac-Spencer says. "This is a bit the same, except that we have much less experience with genomic surveillance than we have with the weather forecast."

To Van Kerkhove, the arrival of B.1.1.7 shows how important it is to follow viral evolution closely. The United Kingdom has one of the most elaborate monitoring systems in the world, she says. "My worry is: How much of this is happening globally, where we don't have sequencing capacity?" Other countries should beef up their efforts, she says. And all countries should do what they can to minimize transmission of SARS-CoV-2 in the months ahead, Van Kerkhove adds. "The more of this virus circulates, the more opportunity it will have to change," she says. "We're playing a very dangerous game here." ■

## COVID-19

# Pfizer's vaccine raises allergy concerns

Polymer in mRNA's "packaging" may cause rare anaphylactic reactions

By **Jon de Vrieze**

**S**evere allergy-like reactions in at least 12 people who received the COVID-19 vaccine produced by Pfizer and BioNTech may be due to a compound in the packaging of the messenger RNA (mRNA) that forms the vaccine's main ingredient, scientists say. A similar mRNA vaccine developed by Moderna also contains the compound, polyethylene glycol (PEG).

PEG has never been used before in vaccines but it is part of many drugs, some of which have occasionally triggered anaphylaxis—a potentially life-threatening reaction that can cause rashes, a plummeting blood pressure, shortness of breath, and a fast heartbeat. Some allergists and immunologists believe a small number of people previously exposed to PEG may have high levels of antibodies against it, putting them at risk of an anaphylactic reaction to the vaccine.

Others are skeptical of the link. Still, the U.S. National Institute of Allergy and Infectious Diseases (NIAID) was concerned enough to convene several meetings last month to discuss the reactions with independent scientists, physicians, representatives of Pfizer and Moderna, and the Food and Drug Administration (FDA). NIAID is also setting up a study in collaboration with FDA to analyze the response to the vaccine in people who have high levels of anti-PEG antibodies or have experienced severe allergic responses to drugs or vaccines before. "Until we know there is truly a PEG story, we need to be very careful in talking about that as a done deal," says Alkis Togias, branch chief of allergy, asthma, and airway biology at NIAID.

Pfizer, too, says it is "actively seeking follow-up." A statement emailed to *Science* noted it already recommends that "appropriate medical treatment and supervision should always be readily available" in case a vaccinee develops anaphylaxis.

Reports about the allergic reactions have created anxiety among potential vaccine recipients. "Allergies in general are so common



At least 12 people suffered an anaphylactic reaction after receiving Pfizer's COVID-19 vaccine.

in the population that this could create a resistance against the vaccines in the population," says Janos Szebeni, an immunologist at Semmelweis University in Budapest, Hungary, who has long studied hypersensitivity reactions to PEG.

Anaphylactic reactions can occur with any vaccine but are extremely rare—about one per 1 million doses. As of 23 December 2020, the United States had seen 10 cases of anaphylaxis among 614,117 people who received the COVID-19 vaccine; the United Kingdom had recorded two. Because the Pfizer and Moderna mRNA vaccines use a new platform, the reactions call for careful scrutiny, says Elizabeth Phillips, a drug hypersensitivity researcher at Vanderbilt University Medical Center who attended an NIAID meeting on 16 December. "This is new."

Clinical trials of the vaccines, which involved tens of thousands of people, did not find serious adverse events caused by the vaccine. But both studies excluded people with a history of allergies to components of the COVID-19 vaccines; Pfizer also excluded those who previously had a severe adverse reaction from any vaccine.

The two COVID-19 vaccines both contain mRNA wrapped in lipid nanoparticles (LNPs) that help carry it to human cells but also act as an adjuvant, a vaccine ingredient that bolsters the immune response. The LNPs are "PEGylated"—chemically attached to PEG molecules that cover the outside of the particles and increase their life span.

PEGs are also used in everyday products such as toothpaste and shampoo as thickeners, solvents, and softeners, and they've been used in laxatives for decades. An increasing number of biopharmaceuticals include PEGylated compounds as well.

The compounds were long thought to be biologically inert, but evidence is growing that they are not. As much as 72% of people have at least some antibodies against PEGs, according to a 2016 study led by Samuel Lai, a pharmaco-engineer at the University of North Carolina, Chapel Hill, presumably as a result of exposure to cosmetics and pharmaceuticals. About 7% have a level that may be high enough to predispose them to anaphylactic reactions, he found. "Some companies have dropped PEGylated products from their pipeline as a result," Lai says. But he notes that the safety record of many PEGylated drugs has persuaded others that "concerns about anti-PEG antibodies are overstated."

Szebeni says the mechanism behind PEG-conjugated anaphylaxis is relatively unknown because it does not involve immunoglobulin E (IgE), the antibody type that causes classical allergic reactions. Instead, PEG triggers two other classes of antibodies, IgM and IgG, involved in a branch of the body's innate immunity called the complement system, which Szebeni has spent decades studying.

In 1999, while working at the Walter Reed Army Institute of Research, Szebeni described a new type of drug-induced reaction he dubbed complement activation-related pseudoallergy (CARPA), a nonspecific immune response to nanoparticle-based medicines, often PEGylated, that are mistakenly recognized by the immune system as viruses. He believes CARPA explains the severe anaphylactoid reactions occasionally caused by some PEGylated drugs, including cancer blockbuster Doxil and pegnivacogin, an experimental coagulant whose phase III trial was halted in 2014

after some participants developed severe allergic responses and one died.

Some scientists believe PEGylated nanoparticles may cause problems through a mechanism other than CARPA. In November, Phillips and colleagues published a paper showing people who suffered an anaphylactic reaction to PEGylated drugs did have IgE antibodies to PEG after all, suggesting those may be involved, rather than IgG and IgM. Other scientists are not convinced PEG is involved at all. "There is a lot of exaggeration when it comes to the risk of PEGs and CARPA," says Moein Moghimi, a nanomedicine researcher at Newcastle University who suspects a more conventional mechanism is causing the reactions. "You are delivering an adjuvant at the injection site to excite the local immune system. It happens that some people get too much excitement, because they have a relatively high number of local immune cells."

Others note the amount of PEG in the mRNA vaccines is orders of magnitude lower than in most PEGylated drugs. And whereas those drugs are often given intravenously, the two COVID-19 vaccines are injected into a muscle, which leads to a delayed exposure and a much lower level of PEG in the blood, where most anti-PEG antibodies are.

Nevertheless, the vaccine companies were aware of the risk. In a 2018 stock market prospectus, Moderna acknowledged the possibility of "reactions to the PEG from some lipids or PEG otherwise associated with the LNP." And in a September paper, BioNTech researchers proposed an alternative to PEG for therapeutic mRNA delivery, noting: "The PEGylation of nanoparticles can also have substantial disadvantages concerning activity and safety." Katalin Karikó, an mRNA vaccine pioneer and senior vice president at BioNTech, says she discussed with Szebeni whether PEG in the vaccine could be an issue. They agreed that given the low amount of lipid and the intramuscular administration, the risk was negligible. Karikó emphasizes that based on what we know so far, the risk is still low. "All vaccines carry some risk. But the benefit of the vaccine outweighs the risk," she says.

Scientists who believe PEG may be the culprit agree, and stress that vaccination should continue. "We need to get vaccinated," Phillips says. "We need to try and curtail this pandemic." But more data on side effects are needed, she adds: "These next couple of weeks in the U.S. are going to be extremely important for defining what to do next." ■

Jop de Vrieze is a science journalist in Amsterdam.



## VOICES OF THE PANDEMIC

# A health economist confronts Kenya's pandemic

Edwine Barasa helps guide the government's response with data—and quiet persistence

By Linda Nordling

In 2007, Kenyan health economist Edwine Barasa had a long layover at Heathrow Airport. A fervent supporter of the London-based soccer club Arsenal, he saw a chance to fulfill a lifelong dream: visiting the club's ancestral stadium in Highbury. Even though he didn't have the right paperwork, he managed to convince an immigration officer to stamp his passport—with a warning that he absolutely had to be back in 12 hours or they would both be in trouble.

The incident speaks to Barasa's tenacity and powers of persuasion, says his boss, Philip Bejon, who directs the Kenya Medical Research Institute (KEMRI)-Wellcome Trust Research Programme. These traits, Bejon says, have served Barasa well in his role as director of the program's office in Nairobi, where he's been a key player in Kenya's response to the COVID-19 pandemic. "Edwine always shows up as an authentic and sincere scientist who convinces his colleagues."

Over the course of the pandemic, Barasa has worked with epidemiologists to reveal the surprisingly small impact of the disease in Kenya—so far. He has advised the country's Ministry of Health on how to allocate its limited resources. And he's been a part of the team guiding KEMRI-Wellcome Trust—a long-standing collaboration between Kenya and the United Kingdom—as it assists the government with testing and viral sequencing, and hosts a Kenyan trial of the COVID-19 vaccine produced by the University of Oxford and AstraZeneca. Barasa believes his field of health economics has much to offer the pandemic response, which entails making life-or-death investment decisions quickly, with limited information.

Until recently, posts like Barasa's—directing internationally funded research partnerships in Africa—were typically held by white Europeans. But Barasa belongs to a generation of young African research leaders now stepping forward. On social media, he has challenged the prevailing powers in the field of "global health"—and the long tradition of researchers from rich northern countries studying poor countries' health problems. He doesn't view

himself as a natural leader, though. "I don't find it comes easy," he says.

Barasa trained as a pharmacist in Kenya before moving to South Africa to do his master's degree and Ph.D. in health economics at the University of Cape Town. Susan Cleary, his postgraduate supervisor and present-day collaborator, describes him as a "superb academic as well as a fine human being—humble, kind, and courageous." Along the way he published five papers while also getting married and becoming a father. He wanted to finish his Ph.D. before



**"We need scientists who are in Africa focusing on African problems."**

Edwine Barasa, Kenya Medical Research Institute-Wellcome Trust

his son was born, but despite completing his degree in a record two-and-a-half years, he didn't quite make that deadline. "I was late by 3 months," he quips.

Back in Kenya, Barasa took leadership of the KEMRI-Wellcome Trust health economic research unit in late 2015. Two years later he was promoted to his current post. He is closely involved in Kenya's efforts to make its health system accessible and affordable for its 53 million citizens, but the pandemic put those reforms on hold.

As coronavirus patients flooded hospitals in Europe early in 2020, Barasa grew increasingly anxious. "I just wondered, if these countries are struggling, what will happen when this pandemic hits the African continent?" In a preprint published on medRxiv in April 2020, Barasa and two colleagues estimated, based on early epidemic modeling, that the demand for intensive care beds in Kenya could outstrip supply by a factor of four.

Luckily, those predictions turned out to be wrong. On 15 April, when some models had predicted Kenya would hit 1000 reported cases, the official tally was just over 200. One month later, when models had predicted more than 10,000 cases, the country had reported only 758. Although official numbers likely undercounted asymptomatic infections, fears that hospitals would be overwhelmed did not come to pass. After Barasa and colleagues modified an existing surveillance system for tracking illness in children to record trends in all-age admissions for severe respiratory symptoms, they found that the number of COVID-19 patients in intensive care in the entire country never exceeded 60 during Kenya's first peak of infections. Even Kenya's limited health system could handle such numbers, he says.

Another clue that the disease was often milder in Kenya came when Barasa and colleagues tested more than 3000 Kenyan blood donors for SARS-CoV-2 antibodies, a sign of previous infection. That work, published in *Science* in November, suggested more than 7% of Nairobi's 4.4 million inhabitants had been exposed to the virus by May. Based on modeling, the researchers concluded infections had peaked in the capital in July with 30% to 50% of the population infected. Yet hospitals were not overwhelmed.

Why the pandemic has played out so differently in Kenya and some other African countries isn't clear. Young people are less susceptible to severe disease, and Kenya's median age of 20 (compared with 47 in Italy) is "pretty much the only factor where there is clear evidence," Barasa says. Other possible factors, such as climate, genetics, or immunity due to previous exposure to other pathogens, haven't yet been fully investigated.

Bejon notes that Barasa has a special

knack for engaging policymakers in the design of research studies, which “has resulted in research becoming more closely attuned to what is needed nationally.” Barasa has a close working relationship with the Ministry of Health, says Kadondi Kasera, a scientist based at the ministry’s Public Health Emergency Operation Center. “We have worked with Edwine and his team to generate a number of policy and evidence briefs that have informed the ministry top management in designing preparedness and response measures,” Kasera says. The work has included calculating the cost of treating a COVID-19 patient in a Kenyan hospital and providing advice on reopening schools.

Barasa has also helped the government target its limited resources. At the start of the pandemic, many feared Kenya didn’t have enough ventilators to keep severely ill patients alive. “That became the narrative,” he recalls. But over time, it became clear to him that instead of spending thousands of dollars on ventilators, which only a few Kenyans would need, it would be wiser to invest in pulse oximeters. These devices measure blood oxygen levels and can be used to determine which COVID-19 patients need supplemental oxygen—insights that benefit many more patients than ventilators.

Kenya’s first pandemic wave proved not to be the tsunami Barasa and others feared. But after peaking in July and August, new cases rose again. Between 12 October and 8 November, the number increased by an average of 34% per week, to just under 1500 cases a day. Since then, both new cases and deaths have been tracking down. Rising immunity may have helped curb both Kenya’s waves, Barasa says, though he won’t say communities are nearing herd immunity. “We don’t know how long immunity lasts,” he adds.

Some Nairobi hospitals may have become overcrowded in October—probably because people from outside the capital were seeking care at the country’s best facilities, Barasa says. Yet with only 22% of Kenya’s population living within a 2-hour walk of a health care facility with intensive care capabilities, the surge in rural cases could bring fresh concerns. And protecting health workers is also proving tricky: Dozens of doctors and nurses have died in the new surge in cases.

Still, with African labs and research centers providing more evidence, governments are now better prepared to face those challenges, Barasa says. “One of the things the pandemic has shown me is that we need local capacity, and we need scientists who are in Africa focusing on African problems.” ■

Linda Nordling is a journalist in Cape Town, South Africa.



Scientists fear oil mapping in an Alaskan refuge could harm polar bears that live along the nearby Beaufort Sea.

## CONSERVATION BIOLOGY

# Alaska oil bid alarms scientists

Mapping plan for Arctic refuge ignores risks, critics say

By Warren Cornwall

**A** plan to crisscross parts of Alaska’s remote Arctic National Wildlife Refuge with earth-shaking machines that help map underground oil formations is drawing criticism from scientists. They warn that such mapping done there decades ago left still-visible scars on the tundra, and they fear the new effort could harm hibernating polar bears.

“This population is already in dire straits,” says biologist Steven Amstrup of Polar Bears International. “Does going in and potentially disrupting them make any sense?”

On 15 December 2020, the federal Bureau of Land Management (BLM) issued a preliminary finding that the mapping would cause no serious damage to the environment. One week earlier, the U.S. Fish and Wildlife Service (FWS) issued a draft permit allowing the work in prime habitat for the bears, which are protected by federal law.

The moves aim to launch oil exploration in the refuge in the waning days of President Donald Trump’s administration. Mapping could start this month, and the federal government plans to auction the first drilling leases on 6 January.

After a long struggle, in 2017 Congress allowed drilling in the refuge’s 600,000-hectare coastal plain for the first time. Now, the proposed mapping of 140,000 hectares by the Kaktovic Iñupiat Corporation, owned by Alaska Natives, is crucial to pinpointing oil reserves. Tremor-generating vehicles, which look like a cross between a bus and a bulldozer, would traverse a rectangular grid 200 meters by 400 meters. Reflections of their seismic waves can reveal what’s below.

Current seismic mapping methods cause

little damage, asserts Shelly Jones, BLM’s Arctic district manager. Jones pointed to similar work in Alaska’s nearby National Petroleum Reserve that, she says, had “no significant impacts, including to wildlife, subsistence, or vegetation.”

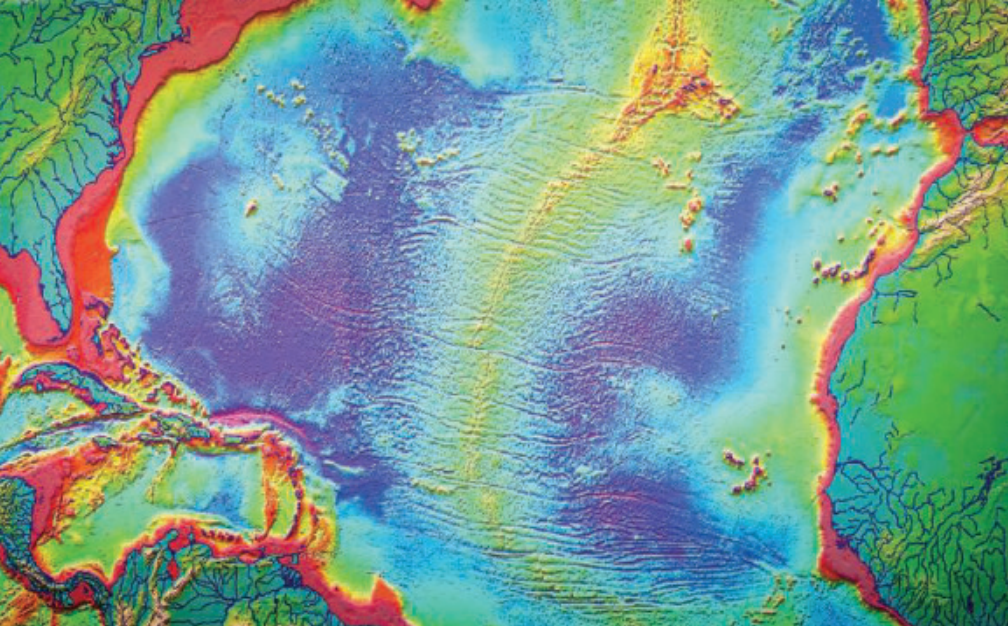
The work would involve as many as 180 people living in bunkhouses dragged across the tundra. To limit impacts, workers will use lighter vehicles with rubber tires when possible, travel atop layers of snow deep enough to protect vegetation, and use heat-sensing cameras to scan for hibernating polar bears buried beneath the snow.

But researchers have done few studies of the effects of such work. During the only other seismic mapping done in the refuge, in the mid-1980s, vehicles crushed vegetation, allowing water to pool and permafrost to thaw, causing areas to sink. In a recent paper, scientists reported 5% of the camp trails still showed damage in 2018. They estimate seismic work across the entire coastal plain could moderately or severely damage 12,200 hectares. The federal agencies have “just kind of developed this storyline that [mapping] impacts are negligible,” says Torre Jorgenson, an ecologist who studied the original seismic work.

Researchers also question FWS’s finding that the work will have a negligible impact on polar bears. The agency estimates the study area will host three polar bear dens. And it predicted that at most three bears might be bothered and no bears would be killed.

But Amstrup fears the machines will disturb or crush hibernating female bears and cubs. The number of bears living at the southern edge of the Beaufort Sea is now about 900, he notes, down 40% since 1980. ■





The extrusion of slabs of crust from midocean ridges slowed 15 million years ago, perhaps cooling Earth.

## GEOPHYSICS

# Slowdown in plate tectonics may have led to ice sheets

Decreased ocean crust production tied to rapid cooling

By **Paul Voosen**

**I**n seafloor trenches around the world, slabs of old ocean crust fall in slow motion into the mantle, while fresh slabs are built at midocean ridges, where magma emerges at the seams between separating tectonic plates. The engine is relentless—but maybe not so steady: Beginning about 15 million years ago, in the late Miocene epoch, ocean crust production declined by a third over 10 million years to a slow pace that pretty much continues to today, says Colleen Dalton, a geophysicist at Brown University who presented the work last month at a virtual meeting of the American Geophysical Union. “It’s a global phenomenon.”

Although previous ocean-spreading records showed hints of a slowdown, nothing suggested such a steep decline, says Clint Conrad, a mantle dynamicist at the University of Oslo who is unaffiliated with the work. The lag was also widespread: Dalton found crust production slowed down or stayed steady at 15 of Earth’s 16 ocean ridges. And its effect on the climate may have been stark, Conrad says. “If you dramatically slow down plate tectonics in such a short time, you can put out a lot less carbon dioxide (CO<sub>2</sub>) gas from volcanism.” The slowdown corresponds to a 10°C drop in temperatures in the late Miocene, when ice sheets began to grow across Antarctica after a long hiatus.

Seafloor spreading is captured in magnetic

zones on the ocean floor. Every million years or so, Earth’s magnetic field flips, and this reversal is frozen in the rocks forged at the midocean ridges. Ship-based observations of the alternating magnetic “stripes” that result as slabs of crust unfurl from the seafloor spreading centers helped give credence to the theory of plate tectonics in the 1960s.

The ridges in the Atlantic and Indian oceans spread slowly, however, which means ships have mapped these stripes with a temporal resolution of only about 10 million years. But geophysicists Charles DeMets of the University of Wisconsin, Madison, and Sergey Merkurjev of Saint Petersburg State University drew on previously unused data from Russian naval ships, which—like those of other nations—tow magnetometers to aid the hunt for enemy submarines. Those data sharpened the resolution in these ocean basins to 1 million years. “And it turns out there are surprising signals hiding in a lot of places that we didn’t know about,” says DeMets, who identified part of the slowdown in his records.

Dalton and her colleagues added to the picture by assembling a complementary high-resolution record for the Pacific Ocean, where seafloor spreading is faster and more complex. With that global view, the slackening immediately became apparent. It appears the deceleration came in two waves, DeMets says: first between 12 million and 13 million years ago in the Pacific and then 7 million years ago in the Atlantic and Indian oceans.

Maybe the subducting slabs stopped tugging as hard on the moving sea floor during this time, Dalton speculates, because they grew thinner or less dense. Or maybe the subduction zones, typically as long as the midocean ridges, shrunk in length, reducing their pull. Another possibility is that the zones changed their orientation, causing the subducting slabs to meet more resistance as they dove into the mantle, which has a kind of natural grain, like wood. Or a slab could have broken off entirely, changing the flow of heat inside the mantle and altering the glide of the tectonic plates overhead, Conrad says. “Even if you change one plate, it affects all the plates.”

By taking volcanic CO<sub>2</sub> emissions tied to today’s ocean crust production and adjusting them for late Miocene speeds, the team found a drop in atmospheric CO<sub>2</sub> that could plausibly explain the global cooling at the time. But Dalton says other explanations are possible—for example, ancient volcanic rocks, newly uplifted out of the ocean to form fresh mountain peaks in places like Indonesia, might have started to soak up more CO<sub>2</sub> (*Science*, 4 January 2019, p. 13). Both mechanisms likely explain some of the drop, says Nicholas Swanson-Hysell, a paleogeographer at the University of California, Berkeley. “But which is more important?”

Beyond lowering CO<sub>2</sub>, the crustal slowdown would have reshaped Earth’s surface. With less seafloor volcanism, the midocean ridges would have been smaller, increasing the capacity of the oceans. Sea levels would have fallen by 22 meters, Dalton calculates, exposing vast new stretches of land. And as the volcanoes went quiet, the planet itself would have grown 5% less efficient at shedding its internal heat, losing some 1.5 terawatts of output—roughly equal to the production of 1500 nuclear power plants. That decline in heat flow wouldn’t have made much difference to atmospheric temperatures, but Dalton says it calls into question reconstructions of Earth’s cooling history that assume constant heat loss across the ages.

Although there’s much to be teased out, it’s clear that, when viewed over relatively short geological time spans, there’s nothing constant about plate tectonics, says Karin Sigloch, a geophysicist at the University of Oxford. “Variation should always be expected.” Slabs break, monster plumes of seafloor magma suddenly erupt—all with huge climatic repercussions for the thin biosphere clinging to life at the surface. Yet they are just burps in a planetary engine that churns away in a deep and hidden underworld. ■

## U.S. SCIENCE FUNDING

# Congress backs research in 2021 spending bill

Modest increases for science complete 4 years of substantial growth despite Trump

By Jeffrey Mervis

Last week, the U.S. Congress completed work on the 2021 federal budget, a new COVID-19 relief package, and its first major energy bill in 13 years. As if moved by the holiday spirit, the lawmakers were as generous as they could be to the U.S. research community.

On 21 December 2020, after months of delays, both houses rushed through a massive annual spending bill that reverses—for the fourth consecutive year—the deep cuts President Donald Trump had proposed for most science agencies. Instead, the \$1.4 trillion package gives budget increases of 3% to the National Institutes of Health (NIH), 2.5% to the National Science Foundation (NSF), 2.3% to NASA science, and 0.4% to the Department of Energy's (DOE's) science office.

Those raises, especially for NIH, disappointed research advocates because they are smaller than in recent years. But they reflect a bipartisan effort by lawmakers to make research budgets a priority as they labored under an agreed-on limit for overall annual spending. And although this year's increases are modest, they cap 4 years of robust growth for science spending under Trump, who many feared would eviscerate research budgets after he took office in January 2017.

As *Science* went to press, Trump hinted he might not sign the bill, potentially complicating its adoption into law. If it is ultimately approved, however, Trump will leave office having seen NIH's annual budget reach \$42.9 billion, a 33% rise over its 2016 level of \$32.3 billion. Similarly, spending by DOE science is now just over \$7 billion, compared with \$5.4 billion in 2017, a boost of 30%. NASA science programs rose by 8% and 11% in 2018 and 2019, respectively, before smaller increases in 2020 and 2021. NSF's budget has grown the least among the four biggest federal science agencies. Even so, a 14% rise since 2017 compares favorably with an overall increase of only 4% during the second term of former President Barack Obama.

How did those budgets do so well under a president widely seen as hostile to science? The answer lies in Congress's authority to set annual spending and its willingness to continually reshape a decade-old budget deal designed to shrink the federal defi-

cit. That deal, negotiated with Obama in 2011, imposed annual caps on civilian and defense budgets that would have trimmed spending by \$1 trillion over 10 years. When Congress ignored the deal in 2013, it required the imposition of mandatory cuts. But lawmakers eventually tired of the caps and negotiated a series of exceptions. The dealmaking continued under Trump, although his initial resistance to the idea led to a 35-day government shutdown as his first year in office ended.

For 2021, Congress gave itself just \$5 billion more to spend on civilian programs,

## A bumpy Trump bump

The deep research cuts proposed each year by President Donald Trump didn't keep these agency budgets from rising during his term.

AGENCY	INCREASE, 2017 TO 2021
National Institutes of Health	25%
National Science Foundation	14%
NASA science	26%
Department of Energy science	30%

including research. That 1% rise meant even the small increases given to research reflect the value that lawmakers place on science. At the same time, says Yvette Seger of the Federation of American Societies for Experimental Biology, “we hope that this is a 1-year anomaly.”

Although spending levels get most of the attention, the eye-glazing 5593 pages of legislation also contain a number of non-financial measures that shape policy at federal agencies. But this year's bill is silent on several hot-button issues affecting the research community.

Lawmakers did not give the Census Bureau a 4-month extension to deliver the results of the 2020 census, something agency officials had previously said they needed to cope with the disruptions caused by the pandemic and a truncated count. Census advocates have vowed to continue to lobby for it in the new Congress.

There is also no mention of a proposal from the chair of the House of Representatives science committee, Representative Eddie Bernice Johnson (D-TX), for a study of systemic racism in U.S. academic

research. Instead, Congress has ordered NASA, NSF, and the National Institute of Standards and Technology to assess the “current racial and cultural makeup” of their workforce and draw up plans to promote “greater acceptance and diversity.”

The \$900 billion COVID-19 relief measure that was attached to the annual spending bill was the fourth emergency measure adopted this year to deal with the devastating impact of the pandemic. But it also fell far short of what research advocates had sought.

Universities wanted \$122 billion to recover from the impact of the pandemic. In addition, they calculated that federal agencies needed at least \$26 billion more to finance research lost or delayed when campuses were shut down in the spring of 2020. The relief bill contains only \$23 billion for higher education, however, and nothing explicitly for bolstering the research enterprise. The shortfall means academic researchers will look to President-elect Joe Biden's incoming administration for help, says Peter McPherson, president of the Association of Public and Land-grant Universities. “We urge lawmakers to view this deal as only a step toward providing more comprehensive relief.”

Biden won't take office until 20 January. But his transition team is already developing both the next relief package and his first budget submission to Congress in February. Those pending legislative initiatives will be the next test of whether science can retain its bipartisan support.

The sweeping energy bill—the first since 2007—includes numerous provisions aimed at boosting science and combating climate change. It sets a 2036 deadline for cutting by 85% U.S. use of planet-warming hydrofluorocarbon chemicals. And it establishes an ambitious goal of spending some \$35 billion over 5 years on energy research. For instance, it calls for the budget of DOE's Advanced Research Projects Agency-Energy, which moves discoveries from the laboratory to the market, to nearly double by 2025 to \$761 million. But those targets are aspirational; Congress must approve such spending through the annual budget process, and often provides amounts that are lower than the targets. ■

With reporting by Jocelyn Kaiser and David Malakoff.





# OPEN ACCESS TAKES FLIGHT

As a new mandate takes effect, researchers and institutions grapple with the trade-offs of making scientific publications free for all

By **Jeffrey Brainard**

In 2018, a group of mostly European funders sent shock waves through the world of scientific publishing by proposing an unprecedented rule: The scientists they funded would be required to make journal articles developed with their support immediately free to read when published.

The new requirement, which takes effect starting this month, seeks to upend decades of tradition in scientific publishing, whereby scientists publish their re-

search in journals for free and publishers make money by charging universities and other institutions for subscriptions. Advocates of the new scheme, called Plan S (the “S” stands for the intended “shock” to the status quo), hope to destroy subscription paywalls and speed scientific progress by allowing findings to be shared more freely. It’s part of a larger shift in scientific communication that began more than 20 years

ago and has recently picked up steam.

Scientists have several ways to comply with Plan S, including by paying publishers a fee to make an article freely available on a journal website, or depositing the article in a free public repository where anyone can download it. The mandate is the first by an international coalition of funders, which now includes 17 agencies and six foundations, including the Wellcome Trust and Howard Hughes Medical Institute, two of the world’s largest funders of biomedical research.



The group, which calls itself Coalition S, has fallen short of its initial aspiration to catalyze a truly international movement, however. Officials in three top producers of scientific papers—China, India, and the United States—have expressed general support for open access, but have not signed on to Plan S. Its mandate for immediate open access will apply to authors who produced only about 6% of the world's papers in 2017, according to an estimate by the Clarivate analytics firm, publisher of the Web of Science database.

Still, there's reason to think Coalition S will make an outside impact, says Johan Rooryck, Coalition S's executive director and a linguist at Leiden University. In 2017, 35% of papers published in *Nature* and 31% of those in *Science* cited at least one coalition member as a funding source. "The people who get [Coalition S] funding are very prominent scientists who put out very visible papers," Rooryck says. "We punch above our weight." In a dramatic sign of that influence, the *Nature* and *Cell* Press

families of journals—stables of high-profile publications—announced in recent weeks that they would allow authors to publish papers outside their paywall, for hefty fees.

Other recent developments point to growing support for open access. In 2017, for the first time, the majority of new papers across all scholarly disciplines, most of them in the sciences, were published open access, according to the Curtin Open Knowledge Initiative. More recently, most major publishers removed paywalls from articles about COVID-19 last year in an attempt to speed development of vaccines and treatments.

Despite these and other signs of momentum, some publishing specialists say Plan S and other open-access measures could be financially stressful and ultimately unsustainable for publishers and the research institutions and authors who foot the bill. As debate continues about just how far and fast the movement will go, *Science* offers this guide for authors readying to plunge in.

## How does open access benefit authors?

Authors who make their work open access may reap benefits, but their magnitude depends partly on what you measure.

One yardstick is a paper's impact. Some studies have reported up to triple the number of citations for open-access articles on average compared with paywalled ones. But authors may be likely to publish their best work open access, which might bring it more citations. A recent analysis that used statistical methods to control for this tendency found a far more modest citation advantage for open access—8%—and only for a minority of "superstar" papers.

Mark McCabe of SKEMA Business School and Christopher Snyder of Dartmouth College studied how citations to articles changed when their journal volumes moved from behind paywalls to entirely open access, and compared them with citations for articles that remained paywalled. For each article in their sample of more than 200,000 papers in ecology and other fields, the researchers accounted for other characteristics that affect citations, such as a paper's age: Newly published papers usually receive a burst of citations at first but fewer later. The modest citation advantage from open access accrued only to high-quality papers, defined as having already garnered 11 or more citations during a 2-year period before the paper became open access, McCabe and Snyder reported in November 2020.

Other studies have found that open-access articles have a larger reach by other measures, including the number of downloads and online views. They also have an edge in Altmetric scores, a composite of an article's mentions on social media and in news stories and policy documents.

These nonscholarly mentions buttress reports that open access enables a broader audience, beyond the core scientific community, to read research findings. In November 2020, Springer Nature and partners released findings from a survey of 6000 visitors to its websites. They reported that an "astonishing" 28% were general users, including patients, teachers, and lawyers. Another 15% worked in industry or medical jobs that required them to read but not publish research.

Even for faculty members who can read subscription-based journals through their institution's libraries, open access could allow quicker access to articles in journals to which the institution doesn't subscribe. Some 57% of academics surveyed said they "almost always" or "frequently" had trouble accessing the full content of Springer Nature's articles.



# How does open access work for authors?



Open access comes in different varieties, or colors, each with its own costs and benefits.

In what's called gold open access, articles carry a license making them freely available on publication. Typically the publisher charges a fee to offset lost subscription revenue and cover the cost of publishing. In recent years, the median paid, after discounts,

was about \$2600, according to a 2020 study by Nina Schönfelder of Bielefeld University. More selective journals, such as *The Lancet Global Health*, have charged up to \$5000. The Nature Research family of journals has set its top open-access fee at €9500 (about \$11,600), and Cell Press will charge \$9900 for its flagship, *Cell*. Some journals are entirely gold open access; other, "hybrid" journals offer authors a choice between free publication behind a paywall or open access for a fee.

A growing number of universities and research institutions, especially in Europe, are striking deals in which they pay a publisher a single fee that covers open-access publishing by their authors and also lets people on their campuses read content that remains behind paywalls. The largest such agreement was reached in 2019 between Springer Nature and 700 German research institutions and libraries. Since the first such deal in 2015, the number grew to 137 in 2020, according to the ESAC Transformative Agreement Registry. However, the deals last year covered publication fees for only 3% of papers produced globally.

A variant called green open access allows authors to avoid publication fees. In this arrangement, authors publish in journals—even

ones that use paywalls instead of charging authors—but also make their article freely available in an online repository. U.S. policy already requires the final, published versions of papers developed with federal funding to be deposited within 12 months in a repository such as the National Institutes of Health's PubMed Central, and many publishers do this automatically. Other authors can use online tools to find repositories. The Directory of Open Access Repositories lists more than 5500 of them.

Publishers typically impose a 6- or 12-month embargo before authors can deposit the final, peer-reviewed version of a paywalled article, but this runs afoul of the Plan S requirement for immediate open access. (The embargo policies of thousands of journals globally are listed in a database called Sherpa/Romeo.) As a compromise, many publishers including the *Science* family of journals allow authors to immediately post a nearly final, peer-reviewed version of a paper in an institutional repository. Plan S accepts this form of green open access, but has added a controversial provision that these accepted manuscripts be licensed for free distribution. Some publishers have complained that this approach threatens their subscription revenues because it could widen free reading of these articles.

Rooryck says Coalition S canvassed major publishers and found none was planning to routinely reject submitted manuscripts funded by Coalition S members because of the prospect that the authors would immediately post them when accepted. A spokesperson for publishing giant Elsevier told *Science* that all its journals will offer authors funded by Coalition S members the option to publish open access for a fee, allowing authors to comply with Plan S without violating embargoes.

# Are publishing fees affordable for authors?



Where a researcher works strongly influences how much money is available for open-access fees. In Europe, institutions used dedicated internal funds to pay fees for 50% of articles their authors published in hybrid journals (those that publish both open-access and subscription content), but in the rest of the world, the figure was only 25%, according

to a 2020 survey of authors by Springer Nature. Authors also tap funders and other sources, including their own personal funds. European scholars reported paying out of their own wallets for just 1% of the articles, compared with 16% in other countries.

In Italy, the Nature group's new €9500 open-access fee has riled some researchers. That figure is "insane, there's no way on Earth to justify that," says Manlio De Domenico, who leads a network science lab at the Bruno Kessler Foundation. The annual research budget for his 10-person lab recently included a total of €8000 for open-access fees for the entire year. "We can spend the money better another way," he says—to pay Ph.D. students and, in normal times, fund travel to conferences and other labs. "To me, the trade-off is clear." (The Nature group says the price reflects its costs to produce such highly selective journals; journals don't normally collect fees for papers they review but don't publish.)

Nor do open-access publication fees hew closely to the laws of demand. One would expect fees to increase with the prestige of the journal, but a recent study by Schönfelder suggests that's not always true. She examined the relationship between fees paid by U.K. funders and the impact factor—a measure based on the average

number of citations per article—of the journals where the papers appeared. She found a strong correlation in journals that published only open-access articles but a weaker one in hybrid journals. Hybrid journals tended to cost more than purely open-access journals, too.

In a paper published last year, Schönfelder suggested her findings reflect the legacy of the subscription prices of large, traditional publishers such as Elsevier and Springer Nature, which publish many hybrid journals. These highly profitable companies with large shares of the publishing market have operated with limited competitive pressure. "If [their] pricing behavior wins through, the open-access transformation will come at a much higher cost than expected today," Schönfelder wrote.

A complete shift to open access could lead publishers to boost publishing fees even further, to try to make up for lost subscription revenues, says Claudio Aspesi, a publishing industry consultant based in Switzerland. Although just over 30% of all papers published in 2019 were paid open access, subscriptions still accounted for more than 90% of publishers' revenues that year, according to Delta Think, a consulting and marketing firm.

Coalition S seeks to exert downward pressure on prices by increasing transparency. When a grantee's research is published, Plan S requires publishers to disclose to funders the basis for their prices, including the cost of services such as proofreading, copy editing, and organizing peer review. Rooryck says the coalition will share the information with authors and libraries, many of which help fund publishing fees. He expects the practice will increase price competition or provide "at a minimum, confidence that some of these prices are fair."

# Who has qualms about open access?



Despite wide acknowledgment by scientists, publishers, librarians, and policymakers of open access' potential benefits, many are reluctant to go all in.

Even in Europe, where the movement for open access has been especially strong, Plan S is unusual. Of 60 funders surveyed in 2019, only 37 had an open-access policy, and only 23 monitored compliance, according to a report prepared for SPARC Europe, a nonprofit that advocates for open access.

Some authors remain hesitant, too. In multiple surveys, authors have ranked open-access publishing below their need to publish in prestigious, high-impact journals to gain tenure and promotion. And they may be wary of a perception among some scientists that journals that carry only gold open-access articles lack rigor. (That view, researchers say, may reflect that such journals are relatively new, which lowers their impact factor.)

A recent study also hints at inequities, finding that established, funded researchers at prestigious institutions are more likely to pay to publish their work open access. Anthony Olejniczak and Molly Wilson of the Academic Analytics Research Center, part of a data firm in Columbus, Ohio, examined the demographics and publishing patterns of more than 180,000 U.S. scholars. Overall, 84% of biological scientists and 66% in the physical and mathematical sciences had authored or co-authored at least one gold open-access paper between 2014 and 2018. Those authors were more likely to have advanced faculty rank and federal grants and to work at one of the 65 leading research universities that belong to the Association of American Universities, Olejniczak and Wilson report in an upcoming paper in *Quantitative Science Studies*.

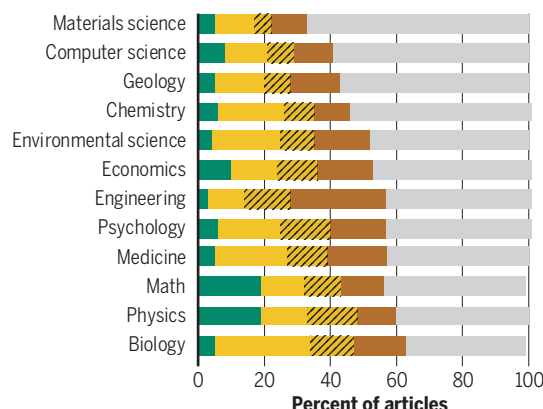
Olejniczak and Wilson hypothesize that scientists who choose to pay for open access not only need financial resources, but also the sense of job security that tenure confers. "This is a good news, bad news story," Olejniczak says. "Open access is thriving, and it's growing." But, he adds, publishers collecting the fees should consider ways to accommodate a wider diversity of authors.

## The many colors of open access

A variety of business models have evolved to support the publication of scientific journal articles that are free to read, and their prevalence differs by field. The Curtin Open Knowledge Initiative performed the analyses using the CrossRef, Microsoft Academic, and Unpaywall bibliometric databases.

### Differences by discipline

The higher rate of gold open access in biology may reflect higher funding levels that cover publication fees. Physics has a long tradition of posting manuscripts in green open-access repositories.



#### Green

Authors or publishers deposit articles in a public repository, where they are free to read. But journal embargoes can delay posting. Numbers shown for green are undercounts because they exclude articles that were also published in other categories of open access (below).

#### Gold

Articles are published with a license making them immediately free to read. Authors or institutions typically pay journals for this service. Gold journals publish only gold articles.

#### Hybrid

Hybrid journals offer gold open-access publication but also publish other articles behind a paywall and continue to charge for subscriptions.

#### Bronze

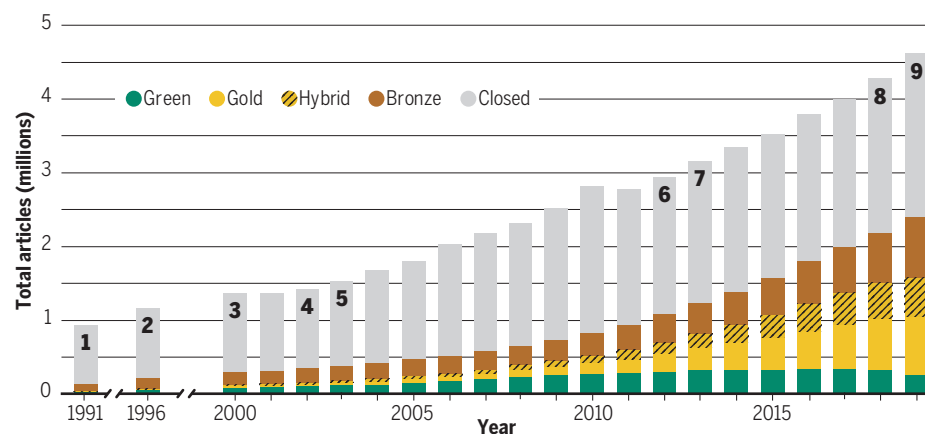
Articles are free to read on publishers' websites, but the papers are not licensed as open access, allowing publishers to place the articles behind paywalls later.

#### Closed

Journals keep articles behind subscription paywalls.

## A gradual opening

In 2017, the percentage of new scientific literature published open access surpassed 50% for the first time. Decisions by authors, publishers, and research funders have helped drive the growth.



**1. 1991** ArXiv, the preprint server that posts papers in physics and other fields, publicly debuts, allowing free online reading of manuscripts.

**2. 1996** *The Journal of Clinical Investigation* becomes the first prominent journal to provide its content free online, as public use of the internet increases.

**3. 2000** BioMed Central, the first open-access, for-profit scientific publisher, starts.

**4. 2002** The Budapest Open Access Initiative defines open-access scholarly articles as allowing the free reuse of the content, with credit to authors.

**5. 2003** The Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities expands on Budapest's terms, calling for research findings and data to be deposited in free, public repositories. The *PLOS* open-access journals are launched.

**6. 2012** More than 2600 scientists vow not to publish in or referee for journals of the publisher Elsevier, in part because of its opposition to a U.S. National Institutes of Health requirement for green open access.

**7. 2013** The White House Office of Science and Technology Policy requires that researchers who publish findings funded by U.S. grants make them open access within 12 months after publication.

**8. 2018** Coalition S, a group of foundations and mostly European funders, announces its Plan S, which requires findings published by its grantees to be immediately open access starting in 2021.

**9. 2019** Springer Nature and German institutions sign the largest "transformative agreement." Such deals allow institutions' authors to publish open access without paying per-article fees.



# Are publishing fees affordable for universities?



One tenet of the open-access movement has been that publishing fees can be funded by redirecting money university libraries currently spend on journal subscriptions—but that assumption faces questions. Although the “transformative” agreements that cover both reading and publishing of articles

have rapidly increased the percentage of articles published open access at some institutions, the details of these deals (like traditional, subscription-only ones) are often secret and have other features that make it difficult to compare bottom-line costs. Comparing costs across institutions is also challenging because these deals usually involve large packages of journals, with the exact lineup varying by institution.

Still, it is clear that making most articles gold open access could wallop the library budgets of research-intensive universities whose scientists publish the most papers. Many institutions that publish little research would save money by dropping subscriptions and letting faculty members read articles for free, analysts say, and publishers would look to recoup the lost revenue through publishing fees.

*Pay It Forward*, a report published by librarians at the University of California (UC) and colleagues in 2016, remains one of the most comprehensive analyses of the impact of these shifts on universities. They calculated what each of UC’s 10 campuses and three comparison institutions would have paid to publish as gold

open access all articles from between 2009 and 2013 that listed one of their faculty members as a corresponding author.

A key finding: At most of the research-intensive institutions studied—such as the UC campuses in Los Angeles and San Francisco and Harvard University—simply redirecting funds from journal subscriptions wouldn’t cover the open-access fees. Those institutions could charge the difference to federal grants, but they would still have to cover fees on papers from studies done without grant funding. Harvard, for example, might have to boost its total library spending by 71%, or nearly \$6 million.

Rich universities like Harvard could potentially tap their huge endowments and copious research funding to cover these costs, but other universities could struggle. U.S. university library budgets have lagged the rate of inflation in higher education for years and now face cuts because of the coronavirus pandemic.

Some researchers interviewed for UC’s study said they were reluctant to spend grant money on open-access publishing fees because they would eat into funds for research. “But in practice, we found [faculty members] are independently spending millions of dollars” from grants on fees, says MacKenzie Smith, university librarian at UC Davis and one of the study’s co-authors. UC is conducting an experiment that limits the universities’ contribution to per-article publication fees in order to encourage faculty members to consider other funding sources and journals with lower fees. “We want to get authors more engaged in the cost aspect of publishing, or at least mindful of it,” Smith says.

## Is open access the future of scientific publishing?



If paying for open-access publication becomes the default route for scientists, and publishers hike prices as expected, many analysts worry publishing will become a luxury that only better funded researchers can afford. That could create a self-reinforcing cycle in which well-funded researchers publish more, potentially attracting more attention—and more funding.

If that comes to pass, it could be especially hard on early-career researchers and authors in the developing world who lack their own grants, and on those in disciplines that traditionally receive less funding, such as math. Although publishers offer waivers for authors, many do not always cover the entire publishing fee or disclose what percentage of requests they grant.

Small, nonprofit societies that currently depend on subscription fees from their journals could also lose out in an open-access world, because the dynamics of the pay-to-publish model tend to favor publishers and journals that produce a high volume of articles, which affords economies of scale.

“I am worried that in the zeal to go that last mile” to make a larger portion of articles open access, “we could end up really hurting the scientific enterprise,” says Sudip Parikh, CEO of AAAS, which publishes the *Science* family of journals. One of them, *Science Advances*, charges an open-access fee of \$4500, whereas the rest operate on the traditional subscription-only model. Parikh says AAAS is considering other options to make papers free to read, but he wasn’t ready to discuss them when *Science* went to press. “I don’t pretend to know the answer yet,” he says. “But it feels like there are other possibilities” besides publishing fees.

One model for sustaining open access without relying on per-article publishing fees comes from Latin America. Brazil and other countries have funded the creation of free open-access journals and article repositories, and the region in 2019 had the world’s highest percentage of scholarly articles available open access, 61%, according to the Curtin Open Knowledge Initiative.

Debate continues about how to control publishing costs. Many advocates for open access say making it more affordable will require a vast shift in the culture of science. In particular, tenure and promotion committees will need to lower their expectations that authors publish in prestigious, costly journals.

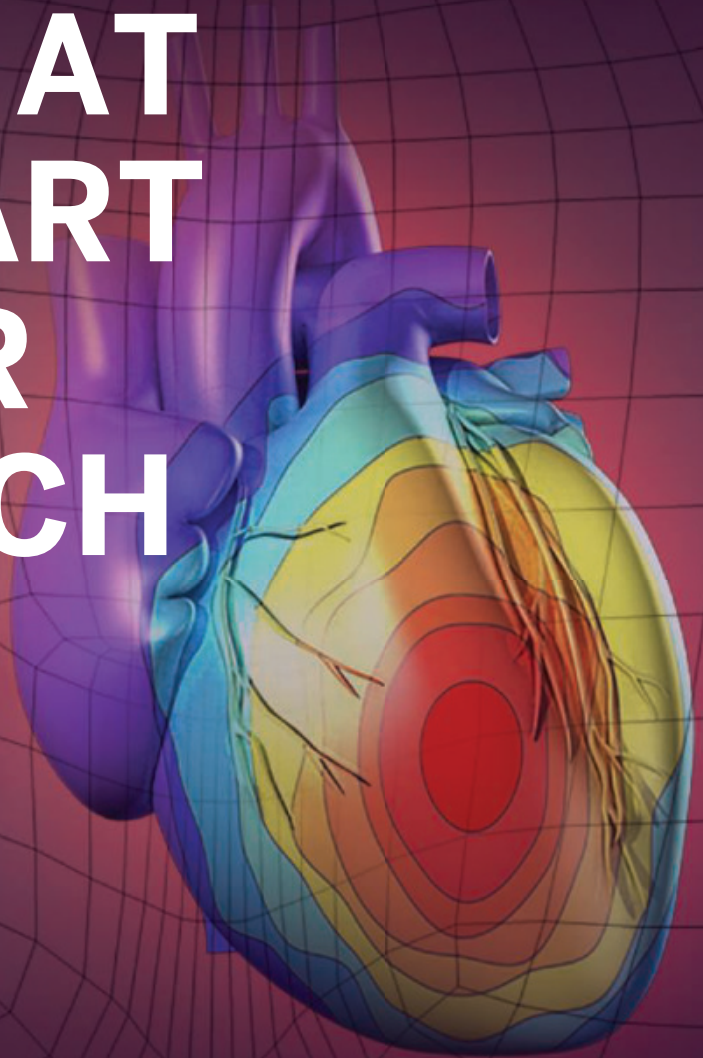
But some argue that even if funders and institutions must cough up more money to help authors publish open access, the potential to accelerate scientific discovery would justify the added cost. The journal publishing industry’s annual revenues of about \$10 billion represent less than 1% of total global spending on R&D—and, in this view, it’s reasonable to divert more of the total to scholarly communications that are essential to making the entire enterprise run.

It’s unlikely, though, that all scientific articles will ever become open access, says Rick Anderson, university librarian at Brigham Young University, who has written extensively about business models for journal publishing. “It just seems to me like the barriers to universal open access are too great,” he says. “Every open-access model solves some problems and creates other problems.”



“What I think is much more likely in the future, almost inevitable, is a fairly diverse landscape of open-access and subscription models,” Anderson adds. “I haven’t yet seen anything that has convinced me that toll [subscription-based] access is going to go away entirely.” ■

# PUT HUMAN HEALTH AT THE HEART OF YOUR RESEARCH

Submit your research:  
**[cts.ScienceMag.org](https://cts.sciencemag.org)**



Science  
Translational  
Medicine  
 AAAS

 Twitter: @ScienceTM  
 Facebook: @ScienceTranslationalMedicine



# INSIGHTS



## NEXTGEN VOICES

### Defining events: 2020 in hindsight

With 2020 finally behind us, we can begin to think about how the historic events that took place will be understood in years to come. To do so, we asked young scientists this question: **What new word or phrase would you add to the dictionary to help scientists explain the events of 2020 to future generations?** Read a selection of the best responses below. Follow NextGen Voices on Twitter with hashtag #NextGenSci. Read previous NextGen Voices survey results at <https://science.sciencemag.org/collection/nextgen-voices>.

#### 2020'd (adjective)

When all of the even slightly negative events in a situation suddenly amplify in magnitude to truly horrendous proportions (e.g., police brutality, political corruption, science skepticism, conspiracy theories, political division). “We are going to need an immediate extraction—things got 2020'd over here.”

**John Protzko**

Department of Psychological & Brain Sciences, University of California, Santa Barbara, CA 93106, USA. Twitter: @JProtzko

#### Algorithmic assurance (noun)

The ability of advanced algorithms to detect and reinforce a person's beliefs in a way that encourages increasingly extremist views. “John did not start out this way, but the algorithmic assurance pushed him further and further until he denied everything from the Moon landing to the existence of the color red.”

**Isaac Z. Tanner**

Vagelos Molecular Life Science Program, University of Pennsylvania, Philadelphia, PA 19104, USA. Email: [itanner@sas.upenn.edu](mailto:itanner@sas.upenn.edu)

#### Biopartisan (adjective)

Of or related to political bias in the interpretation of scientific matters. “In 2020, wearing a face mask to protect against COVID-19 became a biopartisan issue.”

**Morgan Daly Dedyo**

Vagelos Molecular Life Sciences Program, University of Pennsylvania, Philadelphia, PA 19104, USA. Email: [ddedyo@sas.upenn.edu](mailto:ddedyo@sas.upenn.edu)

#### Chaoticatempus (noun)

A state of overwhelming, but transient, chaos. “The promising mRNA vaccine results marked the beginning of the end of the chaoticatempus, also known as the year 2020.” Origin: *Doctor Chaotica* (a Star Trek Voyager character bent on galactic domination and destruction) and Latin *tempus* (time).

**JiaJia Fu**

Whittle School and Studios, Washington, DC 20008, USA. Email: [jjnaturalist@gmail.com](mailto:jjnaturalist@gmail.com)

#### Cryptoshock (noun)

A hidden surprise; unexpected benefits that arise from adversity. “It's a cryptoshock that there are multiple effective SARS-CoV-2 vaccines so soon after identifying the virus.”

**Bhavya Perma**

Vagelos Molecular Life Science Program, University of Pennsylvania, Philadelphia, PA 19104, USA. Email: [bperma23@sas.upenn.edu](mailto:bperma23@sas.upenn.edu)



## Disinforge (verb)

To create information intended to deceive (e.g., related to health, polling, or demographics). “This video looks realistic but is actually a deep fake that was disinforged by a team of experts in order to influence the election.” Origin: English *disinformation* and *forge*.

**Daniel Ari Friedman**

Department of Entomology & Nematology, University of California, Davis, Davis, CA 95616, USA. Email: DanielAriFriedman@gmail.com

## Ecsadtic (adjective)

Rushes of extreme happiness and ecstasy alternating with absolute sadness, causing emotional exhaustion. “When I think about how my family is healthy and my work goals are complete but also how people are suffering from a pandemic and savage fires are consuming our environment, I feel ecsadtic.”

**Ada Gabriela Blidner**

Laboratory of Immunopathology, Institute of Biology and Experimental Medicine—CONICET, C1428 ADN, Buenos Aires, Argentina. Email: adablidner@gmail.com

## Fauci’ing (verb)

To immediately amend or correct statements made by authority figures who misrepresent or overstate findings. “The

graduate student was Fauci’ing at the podium, shoving aside the principal investigator to accurately explain the implications of the results they obtained.”

**Juliet Tegan Johnston**

Department of Physical and Life Sciences, Lawrence Livermore National Lab, Livermore, CA 94550, USA. Twitter @queermsfrizzle

## Fearonomics (noun)

A modern business strategy in which products and services are made available to the masses on the basis of the fears and emotions prevalent in the society. “Adopting a fearonomics business model, the company generated revenue by marketing panaceas geared toward taking advantage of the consumers’ fears of falling victim to the pandemic.”

**Anant Kumar Srivastava**

Asia-Pacific Institute of Management, Jasola Vihar, New Delhi, Delhi 110025, India. Email: anantsrivastava74@gmail.com

## Home-o-static slowficiency (noun)

The ability to work efficiently while attempting to maintain homeostasis under mandatory confinement at home. “In 2020, Fiona was a model of home-o-static slowficiency, publishing five papers and submitting a successful grant application

despite her struggle to maintain balance amidst constant lockdowns.”

**Roland Ruscher and Andreas Kupz**

Australian Institute of Tropical Health and Medicine, James Cook University, Cairns/Smithfield, QLD 4878, Australia. Email: roland.ruscher@jcu.edu.au; andreas.kupz@jcu.edu.au

## Kyrosearch (noun)

The sudden pivot many industries, academics, and government scientists made to confront the pandemic. “Although we’d written grant applications to study broadband network signals, COVID-influenced kyrosearch changed our lab’s focus: We now develop public health interventions using mobile apps.” Origin: Greek *kairos* (opportunity) and English *research*, with spelling reminiscent of Greek *kyrie eleison* (a call for merciful acts).

**Michael A. Tarselli**

TetraScience, Boston, MA 02108, USA. Email: mtarselli@tetrascience.com

## Manusiccrosis (noun)

A condition resulting from frequent hand washing and sanitizing, where the hand becomes irritated, dry, and cracked. “During the COVID pandemic, people developed manusiccrosis from repeatedly washing their hands in an effort to stem the spread of the virus.” Origin: Latin *manus* (hands), *siccum* (dry), and *-osis*, (suffix denoting a process or condition).

**Felicia Beardsley**

Department of Anthropology, University of La Verne, La Verne, CA 91750, USA. Email: fbeardsley@laverne.edu

## Maskonymity (noun)

The inability to recognize the identity or emotions of other people in public places due to the obfuscation of facial features (e.g., by a mask). “Maskonymity makes some people feel isolated but gives others the freedom to pretend they don’t recognize an overly chatty neighbor.”

**Mark Martin Jensen**

Department of Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA. Twitter: @mmjensen3

## Naked noser (noun)

1. A person who wears a mask but leaves their nose uncovered. 2. A person who doesn’t believe in science but is forced (by society or law) to act in accordance with scientific findings. “I wanted to take the subway, but it was full of naked nosers, so I went on foot instead.”

**Nikos Konstantinides**

Department of Biology, New York University, New York, NY 11105, USA. Twitter: @nkonst4



**Nehatha** (noun)

A state of feeling drained of energy to the point of absolute numbness and apathy. “Frontline workers emerged as the nation’s heroes, but some struggled through the pandemic, their isolation and grueling work forcing them into nehatha.”  
Antonym: Sanskrit *hatha* (force).

**Divyansh Agarwal**

Department of Genomics and Computational Biology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA. Twitter: @divyansh\_aga

**Omnidaaichism** (noun)

A universal deterioration or collapse. “Instead of kicking off the new decade with productivity and prosperity, 2020 revealed itself to be an omnidaaichism of our social, political, environmental, and health domains.” Origin: Latin *omnis* (all) and Igbo *daa iché* (fall apart).

**Julia Yuen**

Vagelos Molecular Life Science Program, University of Pennsylvania, Philadelphia, PA 19104, USA. Email: yuenj@sas.upenn.edu

**Pragmaticalopia** (noun)

An inability to see, perceive, or accept facts. “2020 was marked by an increased distrust in science as part of a universal affliction of pragmaticalopia among the denizens of Planet Earth.” Origin: Greek *pragmatiká* (facts), and suffix *-opia* (denoting a visual disorder).

**Suchitra D. Gopinath**

Pediatric Biology Center, Translational Health Science and Technology Institute, NCR Biotech Science Cluster, Faridabad, Haryana, 121001, India. Email: sgopinath@thsti.res.in

**Researchance** (noun)

The process of raising and restoring public confidence in the scientific community, its evidence-based recommendations, and its commitment to improving humanity and the world. “During the COVID-19 pandemic, misinformation, skepticism, and fatigue related to public health measures surged, so the scientific community worked toward researchance through outreach, partnership, and education.”  
Homophone: *resurgence*.

**Michael Tran Duong**

Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. Email: mduong@sas.upenn.edu

**Scienied** (adjective)

Of or related to denying scientific evidence. Pronounced, peddled, and guzzled like cyanide. Embraced by anti-masks and anti-vaxxers. Particularly toxic when mixed with mental delusion and

intentional confusion. “Scienied lies let innocents die.”

**Michael Strong**

Center for Genes, Environment, and Health, University of Colorado, Anschutz Medical Campus, National Jewish Health, Denver, CO 80206, USA. Email: strongm@njhealth.org

**Scifireal** (adjective)

Of or related to science fiction becoming reality in an extremely short period (e.g., 1 year). “All the planes grounded, people wearing masks on the streets, people dying because of lack of medical staff and equipment, vaccines developed in just 9 months—it all felt terrifyingly scifireal.”

**Matúš Soták**

Department of Molecular and Clinical Medicine, Wallenberg Laboratory, Börgeson Lab, Institute of Medicine, University of Gothenburg, Göteborg 41345, Sweden. Twitter: @biomatushiq

**Social networking** (verb)

When someone can’t do their job remotely but can afford to spend the pandemic traveling and posting on social media. “Ugh, don’t you hate having a cousin who has spent the entire summer social networking when you’re trying to balance your job and homeschooling your kids?”

**Katie Burnette**

Department of Evolution, Ecology, and Organismal Biology, University of California, Riverside, Riverside, CA 92521, USA. Email: katiec@ucr.edu

**Tourbivalence** (noun)

Uncertain and contradictory feelings about the impact of COVID-19 on a region’s tourism industry. “Throughout 2020, tourbivalence resulted in frequent debates between Venetians earning a fraction of their normal wages and those enjoying the

peace and beauty of the empty Italian city.”

Origin: English *tourism* and *ambivalence*.

**Samuel Nathan Kirshner**

School of Information Systems and Technology Management, University of New South Wales, Kensington Campus, Sydney, NSW 2052, Australia. Email: s.kirshner@unsw.edu.au

**Trustbot** (noun)

A human fact-checker who works behind the scenes to protect online communities from malicious automated attacks and misleading content. “She is working fewer hours during the COVID-19 lockdown, so to make some extra money she took a part-time job as a trustbot for a popular social networking platform.”

**Athanasia Nikolaou**

Department of Physics, Sapienza University of Rome, Rome, Italy. Email: athanasia.nikolaou@protonmail.com

**Virutopia** (noun)

A city with a population that, to prevent the spread of a deadly virus, turns to an ascetic lifestyle in which people buy only what they need, respect each other, work from home, and reduce their personal contact and entertainment activities. “Life is short in a virutopia, so people live honestly and care for each other.”

**Basant A. Ali**

Energy Material Laboratory, The American University in Cairo, New Cairo, Egypt. Email: basantali@aucegypt.edu

**Zeityrō** (noun)

The collective will to take action to solve a problem, precipitated by a sequence of negative events that promotes profound changes in the accepted way of life. “The extreme weather events, fires, and threats to biodiversity in 2020 led to the zeityrō that allowed people to work together to combat climate change.” Origin: German *zeitgeist* (spirit of the times) and Old Tupi—an Indigenous language of South America—*motyrō* (the union of efforts for the common good).

**Benedito Alves de Oliveira Júnior**

Department of Neuroscience and Behavioral Sciences, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, SP 14049–900, Brazil. Email: benedito.oliveira@usp.br

**Zoomdemic** (noun)

The proliferation of online meetings via the Zoom platform during mandated work-from-home conditions. “Walking on the beach, with the phone off, I finally escape the Zoomdemic and see real people.”

**Elvira Soji**

School of Banking and Finance, University of New South Wales, Sydney, NSW 2052, Australia. Twitter: @esoji

10.1126/science.abg0904

ILLUSTRATION: KATTY HUERTAS



Mammograms are used to diagnose breast cancer, but such procedures have become even more limited owing to COVID-19 restrictions in Africa.

#### VIEWPOINT: COVID-19

## COVID-19 and cancer in Africa

The impacts of COVID-19 present substantial challenges and opportunities in global oncology

By **Beatrice Wiafe Addai**<sup>1,2</sup> and **Wilfred Ngwa**<sup>3</sup>

**T**he COVID-19 pandemic has had a major impact on cancer prevention and control in Africa, with immediate and anticipated long-term ramifications. The pandemic reached Africa when the continent was already struggling to deal with a growing cancer crisis, epitomized by more than 1 million new cancer cases and ~700,000 deaths from cancer per year across Africa (1). The response to COVID-19 immediately exacerbated the challenges in oncology at different levels—including prevention, treatment, and palliative care—and will undoubtedly result in increased late-stage presentation of cancer and a surge in mortality. Meanwhile, efforts to address these challenges have highlighted key opportunities where greater investment could substantially increase access to care and avail global oncology.

At the start of the COVID-19 pandemic,

many African governments responded quickly by shutting their borders, grounding airlines, and limiting travel. These drastic, but necessary, mitigation measures and the diversion of health care resources to address the pandemic resulted in calls by leading hospitals and nonprofit organizations for equal attention to be given to the ongoing cancer epidemic (2). The impact of the COVID-19 measures on oncology was immediate, beginning with cancer prevention, which is particularly important in Africa.

Using Ghana as an example, cancer prevention activities—including awareness, early detection screening, and vaccination—were curtailed. Vital nongovernmental organizations such as Breast Care International (BCI) had to suspend all outreach programs. The literacy rate of ~64% [according to United Nations Educational, Scientific and Cultural Organization (UNESCO)] is relatively low in sub-Saharan Africa (SSA), and cancer is often viewed with superstitions, myths, and misconceptions, which have to be dispelled through outreach programs for education and awareness. Moreover, many African countries lack early detection and screening programs, so it is only through outreach that people can be educated and

clinically examined and/or screened. Such outreach was banned to mitigate COVID-19 spread (2). With many women now at home, BCI started using virtual forms of education—including social media, radio, and other electronic media—to teach women and encourage routine breast self-exams as part of breast cancer screening. The 5-year breast cancer survival rate in SSA is <40% compared with >86% in the United States.

In many countries, screening for cervical cancer, a leading cause of cancer death in Africa (3), was also halted, and medical camps, which can normally screen up to 200 women in a day, could now screen no more than 15 per day because of social distancing guidelines (4). With crucial cancer prevention outreach limited, this will undoubtedly lead to upstaging (that is, diagnosis of more advanced cancers). Because cancer mortality is reduced when it is treated early, this will likely increase cancer mortality. In response to this challenge, some centers are now offering screening and routine human papillomavirus (which causes cervical cancer) immunization when women come to health care facilities for other reasons to limit the number of visits (4). This practice will likely continue

<sup>1</sup>Peace and Love Hospital, Breast Care International, Kumasi, Ghana. <sup>2</sup>Peace and Love Hospital, Breast Care International, Accra, Ghana. <sup>3</sup>Brigham and Women's Hospital, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. Email: drwiae@gmail.com; wngwa@bwh.harvard.edu



beyond the COVID-19 era.

COVID-19 restrictions also reduced cancer diagnostic services, with hospitals postponing diagnostic evaluation or having longer turnaround times because of reallocation of scarce hospital-based resources to tackle COVID-19. In many African countries, biospecimens are transported abroad for diagnostic pathology. During the COVID-19 lockdown, air travel was canceled, so these activities were halted. After the ease of restrictions, some local laboratories have taken the opportunity to improve their services, such that there is now no need to send biospecimens out of the country.

Considering treatment, COVID-19 restrictions resulted in limited access, including reductions in patients traveling to receive treatment and in financial resources to access care. Most people in Africa work in the informal sector, and the restrictions from COVID-19 reduced their income, limiting their ability to pay for treatment. (There is no universal health coverage.) There were considerable reductions in cancer surgery because most clinics were converted to COVID-19 centers. Along with suspension of elective surgical procedures, this has left no time and space for cancer management. Cancer surgeries were triaged into high, medium, and low priorities, and others were canceled.

Radiotherapy (RT) is needed in the treatment of >50% of cancer patients in Africa. RT returned to antiquated two-dimensional (2D) techniques from more effective 3D image-guided RT in some centers owing to limitations in acquiring computed tomography (CT) images, because the scarce CT scanners were being used for COVID-19 patients. Some centers have adopted hypofractionated radiotherapy (HFRT) (5), using higher doses in fewer treatment sessions, owing to limited staff and to reduce the number of times patients need to travel for treatment. Furthermore, patients with concurrent chemoradiotherapy only received RT to minimize possible additional risks of contracting COVID-19 (5).

For chemotherapy, regimens transitioned to less-effective outpatient or oral regimens in many cancer centers, with regional disruptions in supply chains and suspension of blood transfusion services. Some hospitals began using courier service to ensure that drugs were delivered to patients, and larger prescriptions were provided to limit refills. In Africa, >70% of patients present with metastatic cancer, and delays in chemotherapy unfortunately result in deaths. COVID-19 has also severely affected palliative care, with patients being discharged to prioritize COVID-19 patients (4), and many more patients now, distressingly, die in isolation.

Cancer research in Africa has been sub-

stantially scaled down, as seen at the Uganda Cancer Institute (6) and in South Africa, one of the countries with high COVID-19 cases. Fundraising by and for cancer patients has also been severely affected. For example, the Zambian Cancer Society and Women4Cancer in Kenya have documented the difficulties they face in providing care and support during the pandemic owing to redeployment of health care workers to the COVID-19 response (4). Other areas affected by COVID-19 include linking patients with hospital insurance funds to ensure payment and helping patients get alternative accom-

**“...diversion of health care resources to address the pandemic resulted in calls... for equal attention to be given to the ongoing cancer epidemic.”**

modations away from cramped hospital settings. COVID-19 fears and restrictions have added stress to many cancer patients. This is made worse when patients have to delay their cancer treatment. Before COVID-19, the increasing incidence of cancer in Africa had already led to high rates of poor mental health in patients and among family caregivers (7). This is worse during the pandemic (5).

Several opportunities have been highlighted by the effects of the COVID-19 pandemic where greater investment or policy could substantially increase access to cancer care and global oncology. One opportunity is increased adoption of HFRT after the pandemic. Many professional societies and the National Comprehensive Cancer Network recommend that radiation oncology professionals adopt evidence-based treatment guidelines for HFRT to alleviate stress on staff and personnel reductions during the COVID-19 outbreak. Adopting HFRT for cancers with high mortality in Africa, such as breast and prostate cancers, can substantially increase treatment accessibility, reduce treatment cost, and improve patient convenience (8). To ensure safety and maximize the benefits of this approach, increased training for oncology health professionals is needed. Because of the limited number of RT machines in Africa, increased adoption of HFRT is likely to have a more substantial effect on treatment accessibility in Africa than in high-income countries (HICs).

Another area for increased investment that is an important and often underestimated part of the African health care system is in phytomedicine, or the use of plants for

prevention and treatment of diseases, which are used by >80% of cancer patients in Africa (9). With COVID-19 restrictions and populations desperate for treatment, more Africans turned to phytomedicine. Phytomedicine of proven quality, safety, and efficacy is part of the World Health Organization's (WHO's) global health priority of ensuring that all people have access to quality health care. However, phytomedicine use is often driven by anecdotal evidence. Their use delays individuals from seeking health facilities offering conventional treatment, resulting in high rates of advanced stage cancer diagnoses and increased deaths, suffering, and higher cost of treatment. Greater investment is needed in this area, such as supporting implementation of the WHO Traditional Medicine Strategy 2014–2023 (10). This translates to increased investment in science for many reasons, including data on safety and efficacy of phytomedicines, while also identifying candidates with therapeutic potential (11). Research will also drive better policies regulating and integrating evidence-based products and practice into health systems, as appropriate. Furthermore, research can be integrated into education, addressing cultural beliefs around the use of phytomedicine.

Accelerated adoption of information and communication technologies (ICTs) for telemedicine during COVID-19 restrictions has occurred across the world (12). For Africa, which has experienced dramatic gains in ICTs such as mobile phone use and internet in recent years, this presents an important avenue to increase access to health care. Centers in Africa are now using ICTs for remote chemotherapy supervision, symptom management, and palliative care. Where possible, outpatient visits and triage are being shifted to digital consultations to reduce risks of infection. Increasingly, ICTs—such as social media platforms, websites, voice-over messages, and toll-free telecommunication—are being used for oncology services (6). For example, the Cancer Association of South Africa launched tele-oncology services for cancer patients left frustrated by limited access to treatment and support owing to the COVID-19 response (13). There is also increasing adoption of online learning for clinical oncology trainees—for example, in Kenya, Nigeria, Uganda, and Cameroon—including collaboration with faculty from HICs. It is likely that these technologies will continue to be used in the future. Investments in artificial intelligence, as seen in Rwanda to fight COVID-19 (12), could also benefit oncology (14).

There have been differences in the African cancer community's response to COVID-19 compared with that in HICs, which may be attributed to factors such

as limited resources and health care systems. African countries have seen a more consequential impact of resource prioritization away from cancer patients compared with HICs. A welcome difference in response is the growing involvement of the diaspora in telemedicine, such as in virtual tumor boards and e-consultation. This trend is likely to increase and presents an opportunity for Africa to leapfrog into an era of tele-oncology while turning “brain drain” to “brain circulation,” which will strengthen the health system workforce. There is already an emerging vision of building a comprehensive cancer center in the cloud (15) for Africa, accessible from anywhere for consultation, second opinion, follow up, continuous education, and so on, with considerable involvement of the diaspora. During the pandemic, apps have also been developed for the African health care setting that can be extended for use in oncology. For example, the surveillance outbreak response management and analysis system (SORMAS) app used during the recent Ebola outbreak for self-diagnosis and tracing could be adapted for applications in oncology, for example, for collecting symptomatic information and promoting cancer prevention and awareness education. Overall, COVID-19 has been a new challenge with opportunities that can be leveraged in Africa to improve oncology and global health. ■

#### REFERENCES AND NOTES

1. F. Bray *et al.*, *J. Clin.* **68**, 394 (2018).
2. K. Nti, “COVID-19: Don’t lose sight of non-communicable diseases – GNCDA,” 4 April 2020; <https://bit.ly/37IfuOc>.
3. T. R. Rebbeck, *Science* **367**, 27 (2020).
4. Union for International Cancer Control (UICC), Cancer and coronavirus in Africa: the challenges facing volunteer organisations (2020); <https://bit.ly/2K5gS9i>.
5. V. Vanderpuye, M. M. A. Elhassan, H. Simonds, *Lancet Oncol.* **21**, 621 (2020).
6. J. Orem, “Mitigating the impact of COVID-19 in cancer patients: preparedness matters,” 7 April 2020; <https://bit.ly/3IVDVAE>.
7. J. K. Muliira, I. B. Kizza, *Int. J. Africa Nurs. Sci* **11**, 1001667 (2019).
8. O. C. Irabor *et al.*, *JCO Glob. Oncol.* **6**, 667 (2020).
9. M. F. Mahomoodally, *Evid. Based Complement. Alternat. Med.* **2013**, 617459 (2013).
10. World Health Organization (WHO), “WHO traditional medicine strategy 2014–2023” (WHO, 2013).
11. A. Ly, *J. Tumor Med. Prev.* **3**, 555601 (2018).
12. A. Blandford, J. Wesson, R. Amalberti, R. AlHazme, R. Allwihan, *Lancet Glob. Health* **8**, e1364 (2020).
13. N. Mukwevho, “Covid-19 lockdown leaves cancer patients isolated and frustrated,” *Health-E News*, 6 August 2020; <https://bit.ly/33XyX08>.
14. A. Hosny, H. J. W. L. Aerts, *Science* **366**, 955 (2019).
15. W. Ngwa, I. Olver, K. M. Schmeler, *Am. Soc. Clin. Oncol. Educ. Book* **40**, 1 (2020).

#### ACKNOWLEDGMENTS

We thank J. Nkengasong, M. Foote, and oncology health professionals across Africa for helpful discussions.

## CORONAVIRUS

# The puzzle of the COVID-19 pandemic in Africa

More data are needed to understand the determinants of the COVID-19 pandemic across Africa

By Justin M. Maeda and John N. Nkengasong

**T**he COVID-19 pandemic has been puzzling to many public health experts because Africa has reported far fewer cases and deaths from COVID-19 than predicted. As of 22 November 2020, the continent of Africa, comprising 1.3 billion people, had recorded 2,070,953 cases of COVID-19 and 49,728 deaths (1), representing ~3.6% of total global cases (2, 3). Because of the continent’s overstrained and weak health systems, inadequate financing of health care, paucity in human resources, and challenges posed by existing endemic diseases—including HIV, tuberculosis, and malaria—earlier predictions suggested that up to 70 million Africans may be infected with severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) by June, with more than 3 million deaths (4). On page 79 of this issue, Uyoga *et al.* (5) report a serosurvey study (measuring the occurrence of SARS-CoV-2 antibodies) of blood donors in Kenya that suggested that the incidence of SARS-CoV-2 infection is much higher than expected from case numbers.

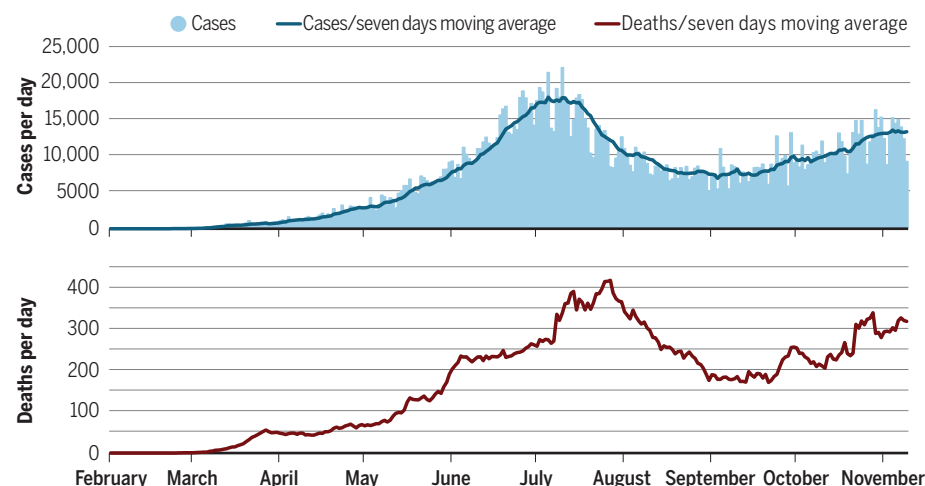
Using blood donor samples as a proxy,

Uyoga *et al.* estimated that SARS-CoV-2 infections occurred in 5.5% of the population in Kisumu, 7.3% in Nairobi, and 8.0% in Mombasa, with an overall average of 4.3%. This translates to ~2.2 million total possible infections compared with the reported 77,585 infections in the country as of 23 November 2020 (1, 3). Similarly, in October 2020, Mozambique reported less than 3000 confirmed cases of COVID-19; however, serosurveys found that 5% of households in the city of Nampula and 2.5% of households in the city of Pemba had been exposed to the virus (6). This suggests that there may be more infections than recorded.

There are several factors that may influence the trajectory of the COVID-19 pandemic in Africa. These include limited testing (which limits detection and isolation, and thus public health measures), a much younger population (and thus fewer severe cases and deaths), climatic differences (which could affect transmission), preexisting immunity, genetic factors, early implementation of public health measures, and timely leadership. Two key aspects that may contribute to our understanding of the pandemic puzzle in Africa include scaling up of

## COVID-19 cases and deaths in Africa

The trend of daily reported cases of COVID-19 for the African continent, February to November 2020, shows the first peak of cases occurred July to August (mostly attributed to the Southern African Region) followed by a second peak, which started in October (mostly attributed to the Northern Region).



10.1126/science.abd1016



testing and use of serosurveys.

One way to unravel the puzzle of SARS-CoV-2 spread is to understand how testing and reporting of cases has occurred. On 14 February 2020, the first cases of COVID-19 were reported in Africa, and by 17 April 2020, the continent had conducted an estimated 330,419 SARS-CoV-2 tests; that is 0.03% of the entire continent's population. In an effort to scale up testing, the Africa Centres for Disease Control and Prevention (Africa CDC) launched the Partnership to Accelerate COVID-19 Testing (PACT) in April 2020. Because of the PACT initiative, testing was scaled up rapidly from ~600,000 per month in April to ~3.5 million per month in November 2020, an increase of ~5.5-fold (7), with 39 of 55 (71%) countries reporting more than 10 tests conducted for every case identified, as recommended by the World Health Organization (WHO) (8). Testing capacity has varied over time, with positivity rate fluctuating between 5 and 15% regardless of the increased testing boost brought by the introduction of PACT.

Therefore, it is clear that testing has been challenging (9, 10), which limits our understanding of the full extent of the spread of SARS-CoV-2 infection in Africa. As such, serosurveys are critical because they can provide data on SARS-CoV-2 infection trends, effects of interventions, demographic characterization, and vaccine effects. Such surveys can also inform on planning for vaccine deployment by providing data to guide prioritization between different populations. They can also aid understanding of the drivers of infection through linking current or previous infection with epidemiological and demographic data. Currently, the continent is facing a challenging phase of the pandemic with an observed second wave of cases (see the figure). More people need to be tested in different localities, including repeated testing over time, so that the patterns and risk factors of viral spread can be understood.

Several serosurveys have been conducted in Africa. The studies differ in methodological approach used: simple random sampling, use of existing sentinel sites, and targeted population (specific subnational unit, pregnant women, blood donors, and people living with HIV). The types of laboratory tests used (rapid tests and enzyme-linked immunosorbent assays) also differed between studies so as to unveil the drivers of infection and disease spread. Given the limited ability to conduct field surveys

(the preferred method) owing to travel restrictions, Uyoga *et al.* investigated blood donors to reveal the pandemic puzzle in Kenya. From these surveys in Africa, seroprevalence of SARS-CoV-2-specific antibodies have ranged from 2.2 to 39% of the population in different settings and countries. However, none of the studies have used a national representative sample.

To ensure a harmonized and standardized method of conducting serosurveys in Africa, the Africa CDC is supporting multinational population-based, age, and gender stratified serosurveys that use standardized protocol and data collection tools (11). The protocol is built on a simplistic model, using point-of-care rapid test for antibody detection of current and previous infection, to ensure feasibility and simplicity while maintaining study quality and credibility of the evidence generated. A similar approach has been applied to national representative cohorts in Brazil and Spain (12, 13).

Across Africa, policy makers are faced with the dilemma of striking a balance between limiting transmission and protecting economies, businesses, and livelihoods. This has created a demand for quality and comprehensive data. Serosurveys could therefore complement existing response strategies. Such surveys should adhere to the following principles: a national representative sample through well-designed sampling strategies that ensure inclusivity of all possible strata within the country; simplicity to guarantee feasibility and quick delivery; optimization of resources for implementation (human, material, and financial) to safeguard the already constrained resources for response; complementarity to already existing surveillance and response data; and the ability to longitudinally track the same aspect of data and information over time to inform adaptive strategies.

Timely leadership and coordination may be a second aspect that explains the pandemic pattern in Africa. The continent reacted in a timely and collective manner once the first cases of SARS-CoV-2 were reported in Egypt on 14 February 2020. Following that, on 22 February 2020, the Africa CDC convened an emergency meeting of all ministers of health at the headquarters of the African Union Commission in Addis Ababa, Ethiopia. The ministers adopted a joint continental strategy that had three goals: limit transmission, limit deaths, and limit social and economic harms and impacts on other endemic diseases, underpinned by the need to coordinate, cooperate, collaborate, and communicate efforts across Africa. In addition, the Africa Taskforce on Coronavirus (AFTCOR) was established to help implement the

strategy and was endorsed by the Bureau of the Heads of State and Governments of the African Union, a validation at the highest level of the continent. This approach helped blunt the early spread of COVID-19.

Therefore, in March 2020, when several countries in Africa began reporting imported cases of COVID-19, there was clarity on the course of action to take. For example, as part of the AFTCOR, the Africa CDC rapidly supported member states to establish diagnostics capacity and expanded testing capacity from two countries in February to more than 43 by the end of March, through competency-based training of member countries at reference centers in Dakar, Senegal, and Johannesburg, South Africa. The coordinated approach ensured harmony in response strategies. For example, the establishment of the African Medical Supply Platform helped to streamline the procurement of response commodities.

The puzzle of the COVID-19 pandemic in Africa can partly be explained by decisive measures taken early to prepare the continent. However, more data are needed to complement what is routinely collected through surveillance and response to understand the different pieces of the puzzle that contribute to the pattern of the pandemic in Africa. Serosurveys and the use of genomic epidemiology can help to better understand disease spread. Further understanding of factors that influence viral pathogenesis and clinical spectrum of disease, and the impacts on endemic infections (HIV, tuberculosis, and malaria), are needed. Efforts to understand attitudes to COVID-19 vaccines are also a priority. ■

## REFERENCES AND NOTES

1. Africa Centres for Disease Control and Prevention. Latest updates on the COVID-19 crisis in Africa (2020); <https://africacdc.org/covid-19>.
2. World Health Organization, WHO coronavirus disease (COVID-19) dashboard (2020); <https://covid19.who.int>.
3. United Nations Department of Economy and Social Affairs, World population prospect 2019 (2019); <https://population.un.org/wpp>.
4. P. G. Walker, C. Whittaker, O. Watson, Report 12 - The global impact of COVID-19 and strategies for mitigation and suppression (2020); <https://bit.ly/37SEg8t>.
5. S. Uyoga *et al.*, *Science* **371**, 79 (2020).
6. A. Frey, Mozambique: 3.79 per cent of Maputo residents exposed to coronavirus (Club of Mozambique, 2020); <https://bit.ly/37Si9J1>.
7. C. D. C. Africa, Outbreak brief 45: Coronavirus disease 2019 (COVID-19) pandemic (2020); <https://bit.ly/3qL0KuE>.
8. Our World in Data—University of Oxford, Test conducted per new case of covid-19; <https://bit.ly/2Kbz4y>.
9. J. Nkengasong, *Nature* **580**, 565 (2020).
10. J. N. Nkengasong, N. Ndembu, A. Tshangela, T. Raji, *Nature* **586**, 197 (2020).
11. C. D. C. Africa, Generic protocol for a population-based, age- and gender- stratified sero- survey study for SARS-CoV-2 (2020); <https://bit.ly/3n5QuuL>.
12. P. C. Hallal *et al.*, *Lancet Glob. Health* **8**, e1390 (2020).
13. M. Pollán *et al.*, *Lancet* **396**, 535 (2020).

Africa Centres for Disease Control and Prevention, Addis Ababa, Ethiopia. Email: justinm@africa-union.org; nkengasongJ@africa-union.org

# RNA-targeted drugs for neuromuscular diseases

Progress with antisense oligonucleotide therapies opens a path for future development

By **Alessandra Ferlini**<sup>1,2</sup>, **Aurelie Goyenvallé**<sup>3</sup>,  
**Francesco Muntoni**<sup>2,4</sup>

**N**euromuscular diseases (NMDs) are common and heterogeneous conditions that either affect skeletal muscle directly, as in muscular dystrophies, or affect motor neurons, peripheral nerves, or neuromuscular junctions. The abundance of skeletal muscle and the size of several of the genes that cause NMDs limit the application of adeno-associated virus-delivered gene therapy (1). Manipulation of RNA to correct mutated transcripts has been used successfully in several NMDs, leading to approved first-generation drugs; next-generation drugs are now in clinical development. But the development of RNA therapies has also been accompanied by several failures, highlighting problems of safety, efficacy, and tissue targeting that need to be overcome.

Manipulation of mutant RNA can be achieved using synthetic antisense oligonucleotides (AONs), which are short, synthetic, single-stranded DNA analogs, either by modulation of splicing (to induce splice switching) or by inactivation. The first splice-switching approach to arrive in the clinic was exon skipping for Duchenne muscular dystrophy (DMD), an X-linked disorder affecting 1 in 5000 live male births. DMD is characterized by progressive muscle weakness and degeneration. It is caused by mutations in the dystrophin (*DMD*) gene that disrupt the open reading frame (ORF) and thus prevent protein production. Dystrophin is a mechanical and signaling scaffold protein linking the actin cytoskeleton and the extracellular matrix, a crucial task for maintaining the integrity of the sarcolemma (the membrane of muscle fiber cells) and avoiding muscle degeneration. In the milder dystrophinopathy variant, Becker muscular dystrophy (BMD), the ORF is preserved, leading to a shortened but functional dystrophin.

The exon-skipping approach to DMD treatment uses AONs that mask pre-messenger RNA (pre-mRNA) splicing sites, resulting in removal of one exon from the mRNA, restoration of the ORF, and production of a BMD-like dystrophin (see the figure).

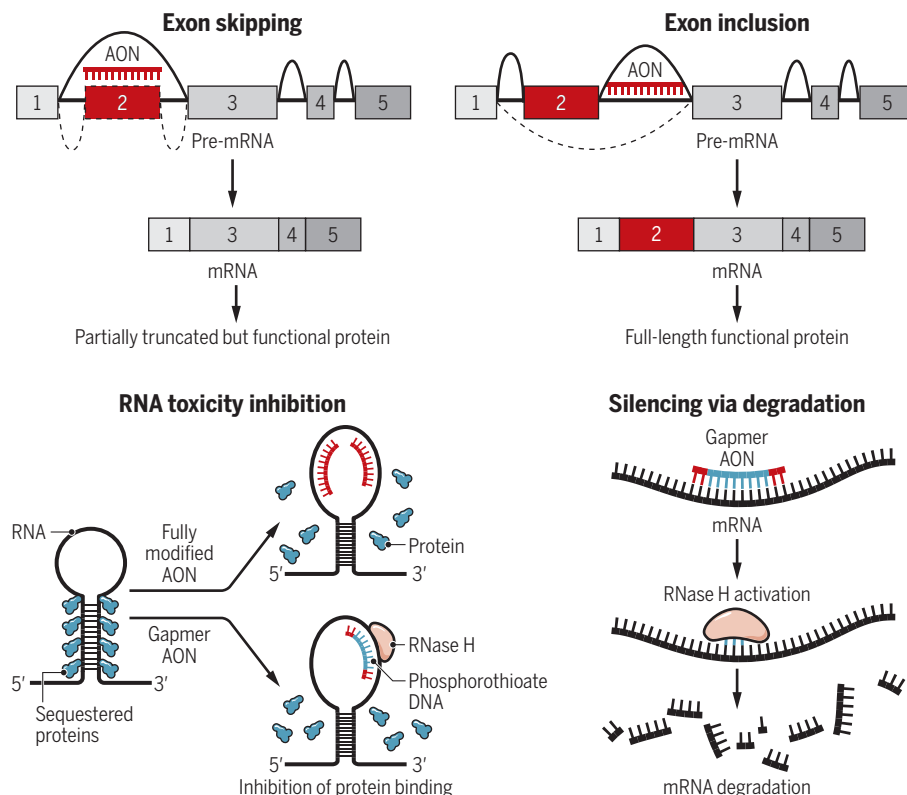
AONs with various modifications to the chemical structure of their “backbone” have been developed to protect them from nuclease activity and to increase their stability and affinity to target RNA. Among them, two chemistries were the first used for *DMD* exon 51 skipping: the charged 2'-*O*-methylphosphorothioate (2'OMe) and the uncharged phosphorodiamidate morpholino oligomer (PMO). Despite encouraging results in initial clinical studies (2), a phase 3 randomized placebo-controlled trial (RCT) with the 2'OMe-modified AON

drisapersen delivered subcutaneously failed to demonstrate significant benefit or clear dystrophin production and was associated with toxicities (such as proteinuria and injection site reactions) that prevented further clinical development (3). By contrast, the neutrally charged PMO eteplirsen, administered intravenously, induced low but significant levels of dystrophin expression and, in a 4-year study, reduced the risk of becoming wheelchair-dependent. In 2016, eteplirsen was granted accelerated approval by the U.S. Food and Drug Administration (FDA), becoming the first approved splice-switching AON and the first approved drug for DMD in the United States.

Other PMO AONs targeting *DMD* exon 53, golodirsen and viltolarsen, were sub-

## Antisense oligonucleotide therapies

Antisense oligonucleotides (AONs) can modulate RNA splicing to induce exon skipping, as in Duchenne muscular dystrophy, where one exon is skipped in the messenger RNA (mRNA) to restore the reading frame and induce Becker muscular dystrophy-like dystrophin. AONs can also force the inclusion of an exon, such as in spinal muscular atrophy (SMA). AONs can be used as steric blockers to inhibit RNA toxicity caused by protein binding. Alternatively, gapmer AONs can induce mRNA degradation through activation of ribonuclease H (RNase H).



<sup>1</sup>Unit of Medical Genetics, Department of Medical Science, University of Ferrara, Ferrara, Italy. <sup>2</sup>Dubowitz Neuromuscular Centre, UCL Great Ormond Street Institute of Child Health, University College London, London, UK. <sup>3</sup>Université Paris-Saclay, UVSQ, Inserm, END-ICAP, Versailles, France. <sup>4</sup>NIHR Great Ormond Street Hospital Biomedical Research Centre, London, UK. Email: f.muntoni@ucl.ac.uk



sequently approved by the FDA (4, 5). Despite these successes, the amount of dystrophin in biopsies from PMO-treated patients remained low, and in 2018 the European Medicines Agency (EMA) gave a negative opinion for eteplirsen, judging that the current risk/benefit balance was unknown because efficacy had not been demonstrated in RCTs, and indicating the use of dystrophin as a surrogate biomarker as premature. To provide conclusive evidence of clinical efficacy, phase 3 RCTs are under way with eteplirsen and golodirsen and also with a PMO targeting exon 45, casimersen.

A limiting step of these AONs is their low efficiency at targeting muscle. The muscle uptake of PMOs is dependent on inflammatory foci associated with dystrophic lesions, as well as the fusion of PMO-loaded monocytes and myoblasts into damaged myofibers. The inefficient targeting of intact muscle fibers also precludes the use of PMOs in conditions with limited muscle damage. Improving AON delivery to muscle is being addressed both with alternative chemistries and new conjugations (6). Stereopure AONs were recently developed to optimize delivery and affinity to pre-mRNA targets. Because the phosphorothioate backbone is chiral, a 20-nucleotide phosphorothioate AON is a mixture of 2<sup>19</sup> different stereoisomers that may not be equally effective. Despite the enhanced potency of the stereopure AON suvodirsen (targeting DMD exon 51) observed in cultured muscle cells, a phase 2 RCT showed no induction of dystrophin expression: Suvodirsen mainly accumulated in the muscle interstitial space, explaining the lack of dystrophin rescue. Clinical development of suvodirsen has been suspended.

Alternative chemistry development includes tricyclo-DNA AONs, a constrained and hydrophobic chemistry allowing higher biodistribution to muscles (7). Moreover, tricyclo-DNA AONs can cross the blood-brain barrier (BBB) after systemic delivery in preclinical studies (7). This might address DMD comorbidities, such as fear and anxiety, caused by dystrophin deficiency in the brain. Conjugation of AONs to cell-penetrating peptides (CPPs) or antibodies targeting specific receptors (such as the transferrin receptor) is also being investigated. CPPs are short cationic and/or amphipathic peptides that facilitate AON translocation across cell membranes and escape from endosomes. Some CPP-conjugated PMO AONs can also cross the BBB in preclinical models (8).

Delivery to the target tissue has been achieved using direct administration in spinal muscular atrophy (SMA). SMA is an

autosomal recessive motor neuron disease with an incidence of 1 in 10,000 live births, caused by inactivating mutations in the survival motor neuron 1 (*SMN1*) gene. SMN is a ubiquitous protein involved in transcriptional regulation and intracellular trafficking, and its deficiency results in selective motor neuron death. SMA patients are classified into four groups of clinical severity; the most common is SMA1 (~60%), in which infants never acquire the ability to sit and typically die before the age of 2 years. A primate-specific gene duplication generated a centromeric variant (*SMN2*) that is present in multiple copies in the population, including SMA patients. *SMN2* allows the production of low amounts of SMN protein, and *SMN2* copy number variation accounts for the differences in SMA clinical severity. The *SMN2* sequence differs from that of *SMN1* by a single-nucleotide polymorphism that weakens its exon 7 splicing enhancer and reduces exon 7 incorporation in the *SMN2* mRNA by 90%, causing the production of an unstable protein.

The splice-switching AON nusinersen enhances *SMN2* exon 7 inclusion, leading to increased production of full-length *SMN2*-derived protein (see the figure). Nusinersen does not cross the BBB and requires direct intrathecal administration [into the cerebrospinal fluid (CSF)]. However, the half-life in CSF is 102 to 111 days, which allows infrequent dosing once a steady state has been achieved. Two phase 3 RCTs have been performed, one in SMA1 and one in the milder SMA2 and SMA3 variants. Nusinersen demonstrated a favorable risk/benefit profile and met the efficacy end points. In SMA1, both survival and acquisition of new motor milestones led to the premature interruption of the RCT, allowing all participants to receive the drug. Similar positive results were obtained in the milder SMA variants. These results prompted FDA and EMA approval (9).

Although the response of symptomatic SMA1 patients is robust and clinically meaningful, many of these children have considerable residual disability due to the advanced stage of the disease before treatment initiation. An ongoing phase 2 open-label trial of nusinersen in presymptomatic infants likely to develop SMA1 or SMA2 found that 88% of those treated achieved walking independently (10). In view of these outstanding outcomes, newborn screening has started in many countries, with important implications for translation (i.e., assessment of the long-term therapeutic impact in presymptomatic patients) and for service provision because of the high prevalence of SMA1.

AONs have been successfully used to si-

lence or down-regulate mRNA. This can be achieved by ribonuclease H (RNase H) activation, which recognizes RNA-DNA hybrids, and subsequent targeted RNA degradation. Degradation of mRNA can be allele-specific (targeting the mutated transcript) or biallelic, so the wild-type transcript is also degraded. As a result of nonspecific binding of AONs, allele-specific transcript silencing is rarely selective, except for alleles carrying repeat expansions. Autosomal dominant diseases caused by repeat expansions arise from the propensity to increase the number of triplets during meiosis due to DNA replication and repair errors (11). Myotonic dystrophy type 1 (DM1) is an autosomal dominant disease affecting the skeletal muscle, heart, and central nervous system. It is caused by CUG repeat expansions in the DM1 protein kinase (*DMPK*) gene that form toxic RNA hairpin structures, which accumulate in the nucleus and induce downstream toxic effects, including sequestration of proteins. AON-mediated silencing can either target and degrade mRNA or inhibit protein binding to the triplets (see the figure). Chimeric AONs, also known as gapmers, induce RNase H-mediated degradation of mRNA containing the CUG repeat expansion, thereby reducing the amount of toxic transcript (12). Optimization of gapmers by constrained ethyl modification increases their binding affinity, further reducing *Dmpk* mRNA by 90% in animal models.

Similarly, AON silencing of expanded hexanucleotide repeats is being pursued in a variant of amyotrophic lateral sclerosis (ALS) due to dominant *C9ORF72* mutations. ALS is a late-onset neurodegenerative disorder characterized by involvement of motor neurons, resulting in progressive paralysis and premature death. ALS is genetically heterogeneous, and *C9ORF72* mutations account for >50% of familial ALS cases. Silencing of *C9ORF72* mRNA containing expanded repeats is being studied in human neuronal cells, as well as in transgenic animals after direct intraventricular administration (13).

A non-allele-selective silencing approach is being pursued for two other autosomal NMDs: centronuclear myopathy (CNM), caused by mutations in dynamin-2 (*DNM2*), and an ALS variant caused by superoxide dismutase 1 (*SOD1*) mutations. *DNM2* is a ubiquitously expressed guanosine triphosphatase (GTPase) mechanosensitive enzyme that is involved in endocytosis, exocytosis, membrane remodeling, and cytoskeletal organization. *DNM2* mutations confer gain of function, and CNM is characterized by the pathological findings of centrally placed nuclei in myocytes, muscle atrophy, and deformed T-tubules. Two different ap-

proaches, including AONs, are being used to target both mutated and wild-type *DNM2* RNA (14). A reduction of total *DNM2* expression in a mouse model improved outcome, providing proof of principle that reduced *DNM2* expression could be therapeutic in CNM. A phase 1 open-label study targeting *DNM2* with a constrained ethyl gapmer in CNM patients has recently started.

Biallelic silencing is also being pursued in SOD1-related ALS. In this condition, dominant missense variants of *SOD1* cause toxic effects in motor neurons by increasing oxidative stress. A phase 1 RCT of an AON targeting *SOD1* through intrathecal administration was well tolerated (15), and a second-generation AON, BIIB067, is entering a phase 3 RCT for SOD1-ALS.

RNA therapies have made impressive progress with the approval of several drugs and further products in the pipeline. Some clinical failures highlight the need to develop alternative chemistries, conjugates, or delivery systems to improve targeted delivery to muscle. The clinical efficacy of next-generation compounds will be enhanced by better understanding of their uptake and intracellular kinetics. For conditions affecting motor neurons, intrathecal delivery efficiently reaches the brain, although chronic administration of these therapies through this route carries a burden for patients; this could be avoided in the future by AONs that cross the BBB. In AON-mediated silencing approaches, biallelic strategies also raise questions about possible haploinsufficiency-related effects and consequent safety profiles. Studies of these next-generation compounds will clarify the extent of clinical benefit and phenotype reversion in these severe conditions. ■

#### REFERENCES AND NOTES

1. T. A. Partridge, *Curr. Opin. Neurol.* **24**, 415 (2011).
2. N. M. Goemans *et al.*, *N. Engl. J. Med.* **364**, 1513 (2011).
3. N. Goemans *et al.*, *Neuromuscul. Disord.* **28**, 4 (2018).
4. Y.-A. Heo, *Drugs* **80**, 329 (2020).
5. H. Komaki *et al.*, *Sci. Transl. Med.* **10**, aan0713 (2018).
6. P. P. Seth *et al.*, *J. Clin. Invest.* **129**, 915 (2019).
7. A. Goyenvall *et al.*, *Nat. Med.* **21**, 270 (2015).
8. S. M. Hammond *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 10962 (2016).
9. A. Aartsma-Rus, *Nucleic Acid Ther.* **27**, 67 (2017).
10. D. C. De Vivo *et al.*, *Neuromuscul. Disord.* **29**, 842 (2019).
11. Y. Liu, S. H. Wilson, *Trends Biochem. Sci.* **37**, 162 (2012).
12. T. M. Wheeler *et al.*, *Nature* **488**, 111 (2012).
13. J. Jiang *et al.*, *Neuron* **90**, 535 (2016).
14. S. Buono *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 11066 (2018).
15. T. M. Miller *et al.*, *Lancet Neurol.* **12**, 435 (2013).

#### ACKNOWLEDGMENTS

A.F. is principal investigator (PI) of Sarepta clinical trials and is a member of the Sarepta European scientific advisory board (SAB). F.M. is PI of Sarepta, Wave Therapeutics, and Biogen clinical trials. F.M. is a SAB member for Biogen, Pfizer, Sarepta, and Dyne Therapeutics. F.M. also receives funding from Sarepta and Biogen.

10.1126/science.aba4515

## MEMBRANES

# Why polyamide reverse-osmosis membranes work so well

## Inhomogeneities in membrane thickness and density promote water transport

By Geoffrey M. Geise

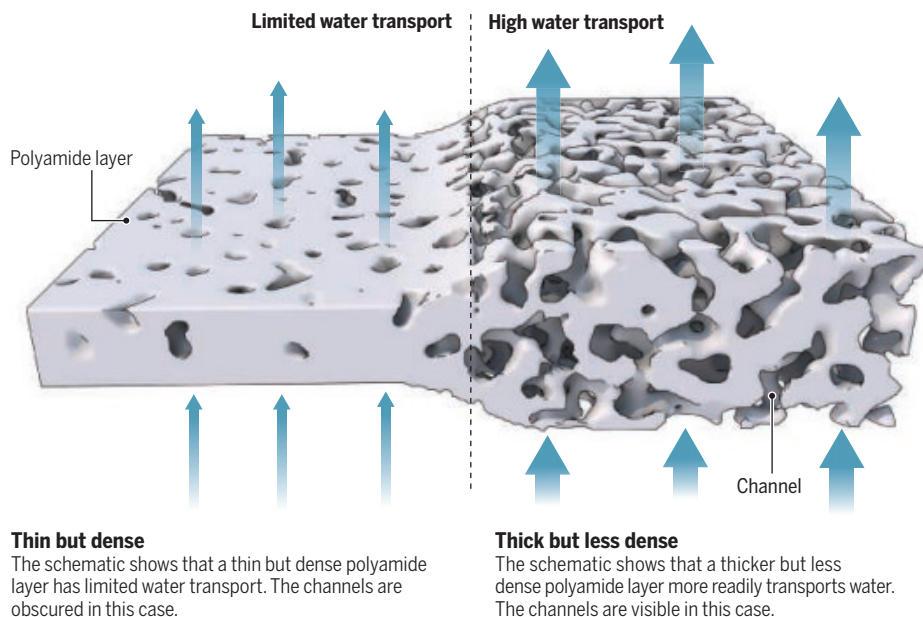
**D**esalination technology leaped forward with the development of interfacially polymerized reverse osmosis (RO) membranes patented by Cadotte in 1981 (1). A huge performance improvement came largely from a selective layer with a self-limiting thickness on the order of 100 nm formed through interfacial polymerization at an oil-water interface on a microporous support (2). Continued membrane property improvements realized during the past four decades have resulted mainly from processing or manufacturing modifications that are often proprietary, rather than by modifying the core chemistry or membrane preparation process (3, 4). The thin, inhomogeneous selective layer has frustrated efforts to satisfactorily characterize polyamide membrane structure and transport properties for decades (5–7). On

page 72 of this issue, Culp *et al.* (8) leverage advances in microscopy and modeling to provide critical insights into structure-property relationships for current state-of-the-art RO membranes that could be used to rationally improve their performance.

Interfacially polymerized RO membranes derive their separation characteristics from a polyamide selective layer. This layer is formed on a microporous support by bringing together an aqueous phase containing an amine monomer (that often is aromatic) and an organic phase containing an aromatic acid chloride monomer (9). Typically, a diamine and trifunctional acid chloride are used to form a highly cross-linked polymer network (10). The reaction between the amine and the acid chloride is very fast (11) and forms the polyamide selective layer within seconds during the manufacturing process, effectively locking in the molecular structure of the membrane. Culp *et al.* directly charac-

## Polyamide membrane density

Culp *et al.* used scanning transmission electron microscopy to image the highly heterogeneous selective polyamide layer of reverse-osmosis membranes. They combined these results with modeling studies to understand variations in water transport.





terize membrane thickness and couple that information with density at the nanoscale. They used scanning transmission electron microscopy with a high-angle annular dark-field detector (HAADF-STEM) to provide a measure of a critical aspect of membrane structure at length scales relevant for describing small-molecule transport. This three-dimensional spatial information revealed where the polymer is (and is not) concentrated, which is important for understanding how water traverses the membrane. Critically, this nanoscale density mapping does not rely on a priori structural assumptions about the membrane material, so the authors could model transport pathways without adjustable parameters. This analysis revealed where water passage occurs in the membrane and, importantly, where water passage is restricted (see the figure).

Conventional macroscopic modeling suggests that passage of water should decrease as the membrane becomes thicker. That is, water passage is restricted as the water is required to navigate a longer path through the membrane. For a material of uniform thickness and density, such behavior would be observed. In RO membranes, however, the nanoscale inhomogeneity in thickness and density observed by Culp *et al.* leads to a situation where water passage can actually increase as the average thickness of the membrane increases if there is a low average density and a narrow density variation. As such, a thick, more homogeneous, and less dense membrane can offer better transport than a thin, dense, heterogeneous one. This behavior, not predicted by conventional modeling, is satisfactorily explained when variations in density and thickness are considered. Notably, bulk average measures of water diffusivity and thickness do not satisfactorily describe the inhomogeneous system caused in part by the nonlinear dependence of water passage on density and thickness.

The approach reported by Culp *et al.* represents a fundamentally different way to characterize RO membranes. The observation that surface-density distributions are important for water passage suggests that this approach may be useful for informing the design of, and directly measuring the impact of, antifouling coatings or other membrane surface treatments that are com-

monly used to prevent biofouling or reduce membrane degradation. Additionally, the microporous supports used in RO membrane manufacturing have been recognized to be critical for determining the ultimate properties of the membrane, and this technique could provide a direct measure of how microporous supports direct the formation of RO membranes during the interfacial polymerization process.

Physical aging or degradation processes can affect long-term material performance of membranes and polymers. For example, when glassy polymers are used in membranes, their properties change over time as the polymer relaxes toward equilibrium (12), and polyamide-based RO membranes are known to degrade in the presence of chlorine (2), which is commonly used as a disinfectant. The approach reported by Culp *et al.* could answer fundamental questions about how RO membranes age, degrade, or both over time.

As Culp *et al.* note, complementary data for salt transport through polyamide materials are still needed to enable a full description of small-molecule passage in RO membranes. Although obtaining such data may not be trivial, the present studies suggest that it should be possible to comprehensively measure and describe small-molecule transport through RO membranes without the need for assumptions about polymer or membrane structure. As such, the HAADF-STEM technique could be a key step toward answering questions that have lingered for decades about how polyamide-based RO membranes function. ■

**“Critically, this nanoscale density mapping does not rely on a priori structural assumptions about the membrane material...”**

#### REFERENCES AND NOTES

1. J. E. Cadotte, Interfacially synthesized reverse osmosis membrane. U.S. Patent 4277344, FilmTec Corporation, Minnetonka, MN (1981).
2. G. M. Geise *et al.*, *J. Polym. Sci., B, Polym. Phys.* **48**, 1685 (2010).
3. Z. Yang, H. Guo, C. Y. Tang, *J. Membr. Sci.* **590**, 117297 (2019).
4. K. P. Lee, T. C. Arnot, D. Mattia, *J. Membr. Sci.* **370**, 1 (2011).
5. Z. Jiang, S. Karan, A. G. Livingston, *Adv. Mater.* **30**, 1705973 (2018).
6. V. Freger, *Adv. Colloid Interface Sci.* **277**, 102107 (2020).
7. D. L. Shaffer, K. E. Feldman, E. P. Chan, G. R. Stafford, C. M. Stafford, *J. Membr. Sci.* **583**, 248 (2019).
8. T. E. Culp *et al.*, *Science* **371**, 72 (2021).
9. W. Xie *et al.*, *J. Membr. Sci.* **403-404**, 152 (2012).
10. E. P. Chan, A. P. Young, J.-H. Lee, J. Y. Chung, C. M. Stafford, *J. Polym. Sci., B, Polym. Phys.* **51**, 385 (2013).
11. T. D. Matthews, H. Yan, D. G. Cahill, O. Coronell, B. J. Mariñas, *J. Membr. Sci.* **429**, 71 (2013).
12. X.-X. Low, P. M. Budd, N. B. McKeown, D. A. Patterson, *Chem. Rev.* **118**, 5871 (2018).

Department of Chemical Engineering, University of Virginia, Charlottesville, VA 22904, USA. Email: geise@virginia.edu

10.1126/science.abe9741

#### STEM CELLS

## Detecting oxygen changes in the lungs

Lung airway basal stem cells directly sense changes in oxygenation, driving lung regeneration

By William Zacharias

**T**he lungs exist in a distinct environment of constantly changing oxygen concentrations. Resultant tissue hyperoxia and hypoxia cause substantial organismal stress, and cells have evolved specific pathways to respond to such changes. This is a particularly acute challenge to the lungs, where differences in the partial pressure of oxygen ( $P_{O_2}$ ) in the circulation and inspired air must be integrated to ensure appropriate cellular responses and maintain tissue homeostasis. On page 52 of this issue, Shivaraju *et al.* (1) characterize how airway basal stem cells sense and respond to hypoxia (low  $O_2$  tension), driving expansion of solitary neuroendocrine (NE) cells that generate paracrine signals to improve airway regeneration in mice. These data demonstrate the complexity of the mechanisms underlying how the lungs respond to changes in tissue oxygenation after injury.

The  $P_{O_2}$  of ambient air is ~160 mmHg and decreases during its transit from the nasal orifice to the air sacs (alveoli) in the lungs, reaching 100 mmHg in the distal lung. The alveolar-arterial gradient and mixture of oxygenated and deoxygenated blood in alveolar venules lead to a systemic arterial  $P_{O_2}$  of 95 to 100 mmHg that is delivered to end organs (2). Assuming intact cellular metabolism, returning venous blood ranges in  $P_{O_2}$  from 40 to 50 mmHg. Under conditions of reduced inspired oxygen concentration, such as high elevation,  $P_{O_2}$  is reduced throughout the lungs, causing regional vasoconstriction of pulmonary vasculature to match ventilation and perfusion within the most-oxygenated alveoli, a phenomenon called hypoxemic pulmonary vasoconstriction. As arterial  $P_{O_2}$  falls, specialized oxygen-sensing cells in the arterial carotid body and airway

NE cells in specialized lung neuroepithelial bodies (NEBs) (3) stimulate respiratory drive, and adrenal chromaffin cells secrete catecholamines to activate the sympathetic nervous system to increase cardiac output and respiratory rate. These physiologic responses improve oxygenation during exercise and under conditions of reduced respiration but have limited impact in the presence of low inspired  $PO_2$ .

Shivaraju *et al.* examine the response of the airway in mice after prolonged exposure to 8% inspired  $O_2$  ( $PO_2 \sim 60$  mmHg), which causes widespread systemic hypoxia. In hypoxic cells, multiple cellular mechanisms are activated (4), affecting transcription, proteostasis, metabolism, and energy production (5). Mitochondria are key sensors of hypoxia, producing reactive oxygen species (ROS) during electron transport, which accumulate during cellular hypoxia. In the lungs, ROS accumulation increases cellular  $Ca^{2+}$  flux in vascular smooth muscle cells and causes hypoxic pulmonary vasoconstriction (6). Cell-specific oxygen-sensitive enzymes of the reduced nicotinamide adenine dinucleotide phosphate (NADPH) oxidase family (NOX) function in diverse processes in the lung epithelium, including barrier function, apoptosis, and proliferation. NOX2 has been directly implicated in oxygen sensing by NEBs (7).

Hypoxia also inactivates prolyl hydroxylase domain proteins (PHDs) to activate hypoxia-inducible factor (HIF)-mediated transcription. During normoxia, PHDs bind and hydroxylate proline moieties on the HIF1 $\alpha$  subunit, promoting its ubiquitination and degradation. Under hypoxic conditions, HIF1 $\alpha$  is stabilized, then heterodimerizes with HIF1 $\beta$  to form an active transcription factor that activates hypoxia-induced gene programs that affect the balance of cellular proliferation and apoptosis, differentiation, angiogenesis, and motility (8). These oxygen sensors can be activated in many cells during severe arterial hypoxia, when tissue  $PO_2$  is less than 25 mmHg. When hypoxia is prolonged, necrotic or apoptotic cell death occurs, leading to tissue injury.

The study by Shivaraju *et al.* adds several layers to our understanding of the cellular response to prolonged hypoxia in the airway. They show that solitary NE cells (scattered in the airway rather than in NEBs) expand after hypoxia, predominantly arising from basal cells. Notably, deletion of all three PHD proteins in basal cells increased the solitary NE cell population. These solitary NE cells secreted calcitonin gene-related peptide I (CGRP), which acts on

other organs may also respond to arterial hypoxia. Given the importance of cellular metabolism in progenitor biology (11), it is possible that hypoxia sensing may be a common property of tissue progenitors, leading to tissue-specific protective or regenerative programs during systemic hypoxia.

How can the findings of Shivaraju *et al.* be integrated with our knowledge of respiratory physiology at a clinical level? During lung injury, patients are frequently

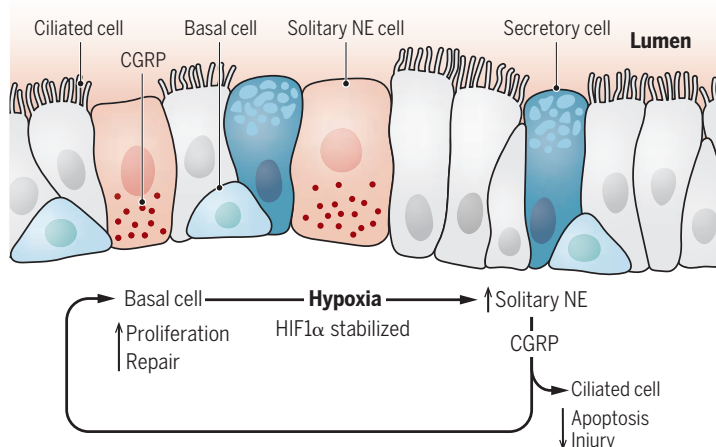
exposed to supplemental oxygen, sometimes with the additional stress of mechanical ventilation. High inspired  $O_2$  concentrations cause lung toxicity, largely through ROS-mediated inflammation, and targeting lower arterial  $PO_2$  may improve mortality in severe acute lung injury (12). The findings of Shivaraju *et al.* imply that NE expansion in the airway is a beneficial adaptation to hypoxia and suggest that artificially high oxygenation targets may impair the regenerative response to injury. Thus, oxygen sensing by basal cells may function as a rheostat to adjust regeneration according to the amount of airway hypoxia.

Conversely, higher arterial  $PO_2$  may improve cognitive outcomes in patients with critical lung injury (13), emphasizing that what is “good” for the lung

may not be ideal for other organs. This provocative and clinically relevant dichotomy points to a need for further studies to specifically understand the oxygen-sensing capacity and effects of hypoxia on cells throughout the body to define the “correct” oxygenation target for critically ill patients. ■

## Airway response to hypoxia

In response to hypoxia, solitary neuroendocrine (NE) cells in the airway linings secrete calcitonin gene-related peptide I (CGRP) in a hypoxia-inducible factor (HIF)-dependent manner. This has paracrine effects on ciliated and basal cells to protect the airways from injury and to promote repair.



other airway cells to promote repair (see the figure). CGRP receptors were broadly up-regulated in the hypoxic airway epithelium, priming them for regeneration. These data demonstrate that hypoxia sensing by the PHD-HIF pathway in basal cells causes NE cell expansion and secretion of CGRP, which serves as a key regulator of regeneration. NE hyperplasia accompanies many chronic hypoxic childhood and adult lung diseases, especially fibrotic disease (9). Neuroendocrine hyperplasia of infancy, a pediatric interstitial lung disease presenting with hypoxia and respiratory distress, is characterized by excess NE cells (10). Could such NE expansion be due to physiological or pathological activation of basal cell oxygen sensing or CGRP signaling in the airway?

Because lung cells are exposed to multiple different oxygen tensions at once, it is unclear exactly what  $PO_2$  basal cells are sensing. Basal cells are localized to the base of the epithelial layer, close to vasculature with limited access to the airway lumen. Perhaps basal cells primarily sense arterial oxygenation. This suggests that progenitor cells in

## REFERENCES AND NOTES

1. M. Shivaraju *et al.*, *Science* **371**, 52 (2021).
2. R. L. Riley, A. Cournand, *J. Appl. Physiol.* **1**, 825 (1949).
3. P. J. Kemp *et al.*, *Am. J. Respir. Crit. Care Med.* **166** (suppl. 1), S17 (2002).
4. J. P. T. Ward, *Biochim. Biophys. Acta Bioenerg.* **1777**, 1 (2008).
5. P. Lee *et al.*, *Nat. Rev. Mol. Cell Biol.* **21**, 268 (2020).
6. P. I. Aaronson *et al.*, *J. Physiol.* **570**, 53 (2006).
7. X. W. Fu *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 4374 (2000).
8. V. L. Dengler *et al.*, *Crit. Rev. Biochem. Mol. Biol.* **49**, 1 (2014).
9. S. Shyue *et al.*, *Pathology* **50**, 699 (2018).
10. M. G. O'Connor *et al.*, *Ann. Am. Thorac. Soc.* **12**, 1730 (2015).
11. N. Shyh-Chang, H.-H. Ng, *Genes Dev.* **31**, 336 (2017).
12. N. R. Aggarwal, R. G. Brower, *Ann. Am. Thorac. Soc.* **11**, 1449 (2014).
13. M. E. Mikkelsen *et al.*, *Ann. Am. Thorac. Soc.* **11**, 613 (2014).

Departments of Pediatrics and Medicine, Divisions of Pulmonary Biology and Pulmonary and Critical Care Medicine, University of Cincinnati School of Medicine, Cincinnati Children's Hospital, Cincinnati, OH, USA.  
Email: william.zacharias@cchmc.org



## POLICY FORUM

## GEOSCIENCE

# Mapping the global threat of land subsidence

Nineteen percent of the global population may face a high probability of subsidence

By Gerardo Herrera-García<sup>1,2,3</sup>, Pablo Ezquerro<sup>1,4</sup>, Roberto Tomás<sup>2,5</sup>, Marta Béjar-Pizarro<sup>1\*\*</sup>, Juan López-Vinielles<sup>1,6</sup>, Mauro Rossi<sup>7</sup>, Rosa M. Mateos<sup>1,3</sup>, Dora Carreón-Freyre<sup>2,8</sup>, John Lambert<sup>2,9</sup>, Pietro Teatini<sup>2,10</sup>, Enrique Cabral-Cano<sup>2,11</sup>, Gilles Erkens<sup>2,12,13</sup>, Devin Galloway<sup>2,14</sup>, Wei-Chia Hung<sup>2,15</sup>, Najeebullah Kakar<sup>2,16</sup>, Michelle Sneed<sup>2,17</sup>, Luigi Tosi<sup>2,18</sup>, Hanmei Wang<sup>2,19</sup>, Shujun Ye<sup>2,20</sup>

**S**ubsidence, the lowering of Earth's land surface, is a potentially destructive hazard that can be caused by a wide range of natural or anthropogenic triggers but mainly results from solid or fluid mobilization underground. Subsidence due to groundwater depletion (1) is a slow and gradual process that develops on large time scales (months to years), producing progressive loss of land elevation (centimeters to decimeters per year) typically over very large areas (tens to thousands of square kilometers) and variably affects urban and agricultural areas worldwide. Subsidence permanently reduces aquifer-system storage capacity, causes earth fissures, damages buildings and civil infrastructure, and increases flood susceptibility and risk. During the next decades, global population and economic growth will continue to increase groundwater demand and accompanying groundwater depletion (2) and, when exacerbated by droughts (3), will probably increase land subsidence occurrence and related damages or impacts. To raise awareness and inform decision-making, we evaluate potential global subsidence due to groundwater depletion, a key first step toward formulating effective land-subsidence policies that are lacking in most countries worldwide.

A large-scale systematic literature review reveals that during the past century, land subsidence due to groundwater depletion occurred at 200 locations in 34 countries [see supplementary materials (SM)]. However, subsidence extent is only known for one-third of these records, information on the impacts is scarce, and mitigation measures were implemented only in a few locations. In China, widespread subsidence affects cities developed in the main sedimentary basins. In Indonesia, coastal subsidence in Jakarta is so severe that government authorities are planning to move the capital to the island of

Borneo. In Japan, subsidence affected several cities during the 20th century, including more than 4 m of subsidence in Tokyo, before groundwater management practices mitigated further subsidence. Iran currently hosts some of the fastest-sinking cities in the world (25 cm year<sup>-1</sup>) because of unregulated groundwater pumping. In Europe, the greatest impact of subsidence occurs in the Netherlands, where subsidence is primarily responsible for placing 25% of the country below the mean sea level and increasing the flooding risk. Subsidence in the Po River Plain in Italy started during the second half of the 20th century and currently threatens 30% of the Italian population, contributing to recurrent coastal flooding during extreme high tides in Venice. In North America, intense groundwater depletion triggers subsidence from California's Central Valley, with as much as 9 m of subsidence in the past century, to the Atlantic and Gulf of Mexico coastal plains in the United States, where subsidence is increasing flooding risk. In México, subsidence rates are among the highest worldwide (as much as 30 cm year<sup>-1</sup>), affecting small structurally controlled intermontane basins where the main urban centers developed, causing an important but unaccounted economic impact.

Spatial analysis of subsidence locations identified in our global database (see SM) reveals that subsidence has preferentially occurred in very flat areas where unconsolidated sediments accumulated in alluvial basins or coastal plains, and where urban or agricultural areas developed in temperate or arid climates characterized by prolonged dry periods. Land subsidence has generally occurred in water-stressed basins, where the combination of groundwater withdrawal and natural groundwater discharge outpaced groundwater recharge, resulting in groundwater storage losses, groundwater depletion, and compaction of susceptible aquifer systems. In the affected basins, land subsidence mainly occurred in highly populated areas, with half of documented occurrences in ar-

reas susceptible to flooding. In coastal zones, the combined effects of absolute sea-level rise and land subsidence contribute to relative sea-level rise (4). The contribution from land subsidence may exceed the contribution from absolute sea-level rise by a factor of 10 or more and could be especially critical for 21% of the geographic locations identified in our database, where land elevation is less than 1 m above the mean sea level.

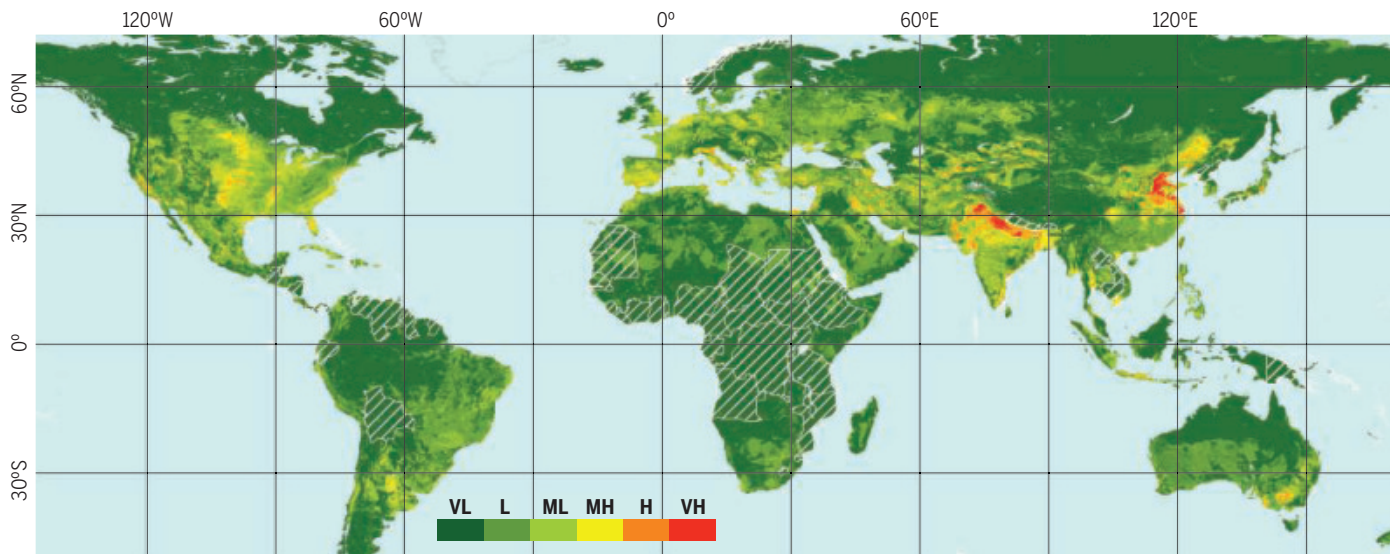
On the basis of the spatial analysis findings, a global model is proposed to combine the main variables influencing subsidence to identify environmental settings favoring land subsidence and the anthropogenic factors leading to groundwater depletion (see SM). Statistical analyses of lithology, land-surface slope, land cover, and Koppen-Geiger climate classes are used to predict global subsidence susceptibility at a spatial resolution of 1 km<sup>2</sup>. The probability of groundwater depletion is estimated by identifying urban and irrigated areas suffering water stress and where groundwater demand is high.

The analyses do not consider subsidence magnitude and rate, owing to the lack of this information at a global scale. Hence, the combination of subsidence susceptibility and the probability of groundwater depletion is used to predict a "proxy" of subsidence hazard, which permits identification of exposed areas where the probability of land subsidence occurrence is high or very high. Even though these results do not necessarily translate to direct impacts or damages, they are useful for identifying potential subsidence areas where further local-scale analysis is necessary. The comparison of our model predictions with an independent validation dataset reveals a 94% capability to distinguish between subsidence and nonsubsidence areas, according to the value of the area under the receiver operating characteristic curve (see SM). The global exposure to potential subsidence is evaluated by calculating the number of inhabitants living in potential subsidence areas, i.e., subsidence hazard proxy, and the equivalent gross domestic product (GDP). This "proxy" of exposed assets is calculated assuming that GDP per capita is homogeneous within each country. Finally, the evolution of potential global subsidence and the related exposure is pre-

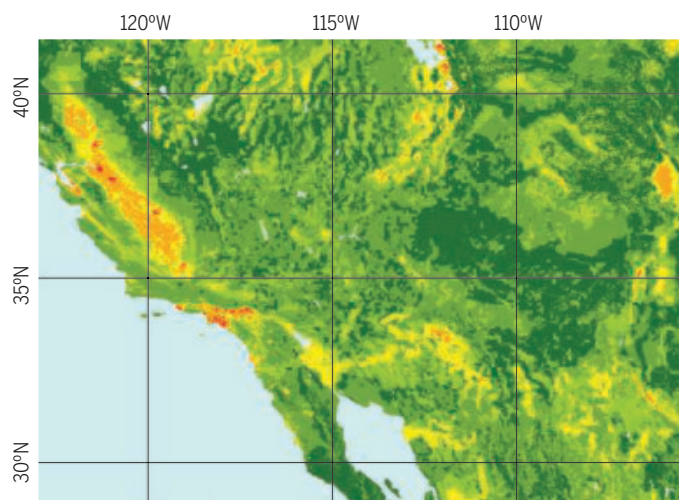
See supplementary materials for author affiliations.  
Email: g.herrera@igme.es

## Potential global subsidence

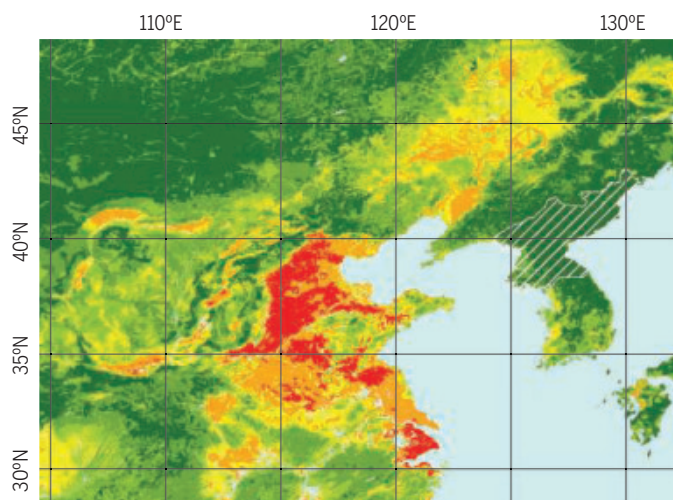
The color scale indicates the probability intervals classified from very low (VL) to very high (VH), for every 30-arcsec resolution pixel (1 km by 1 km at the Equator). The white hatched polygons indicate countries where groundwater data is unavailable, and the potential subsidence only includes information on the susceptibility. See maps of other regions in supplementary materials.



**North America**



**East Asia**



dicted for 2040 for a global change scenario based on steady population growth and increasing greenhouse gas emissions (Shared Socioeconomic Pathways 2, Representative Concentration Pathway 8.5), which accounts for the greatest sea-level rise projections.

Our results suggest that potential subsidence threatens 12 million km<sup>2</sup> (8%) of the global land surface with a probability greater than 50% (MH to VH in the figure). Potential subsidence areas are concentrated in and near densely urban and irrigated areas with high water stress and high groundwater demand, overlying some of the largest and most depleted aquifer systems (5) in Asia (e.g., North China Plain) and North America (e.g., Gulf of Mexico coastal plain);

coastal and river delta areas worldwide (e.g., Vietnam, Egypt, or the Netherlands); and inland sedimentary basins of México, Iran, and the Mediterranean countries. Potential subsidence is lower in Africa, Australia, and South America, owing to the lower groundwater depletion (6). In central Africa, potential subsidence only includes information on the susceptibility, as groundwater depletion is unknown. In this region, subsidence susceptibility (see fig. S6) could be useful to prevent subsidence impacts on developing cities that during the next decades could rely more on the available groundwater resources.

To evaluate the exposure to potential subsidence, we focus on areas where the potential subsidence probability is high or very

high (see the figure). The cumulative potential subsidence area amounts to 2.2 million km<sup>2</sup>, or 1.6% of the land; includes 1.2 billion inhabitants, or 19% of the global population; and has an exposed GDP of US\$ 8.19 trillion, or 12% of the global GDP. High-income countries account for 62% of the global exposed GDP but only 11% of the global exposed population, whereas low-income countries account for 54% of the global exposed population and 12% of the global exposed GDP. It is expected that the capability of low-income countries to implement the political, regulatory, and socioeconomic measures necessary to prevent and mitigate subsidence impact will be less than that for high-income countries.



Potential subsidence threatens 484 million inhabitants living in flood-prone areas, 75% of whom live in fluvial areas and 25% of whom live near the coast. This number of threatened inhabitants corresponds to 50% of the global population exposed to flooding hazards according to previous estimates (7), demonstrating the importance of considering potential subsidence in global flooding risk analyses.

Most of the global population exposed to potential subsidence live in Asia (86%), which is about 10 times the combined exposed population of North America and Europe (9%). The results indicate that 97% of the exposed global population is concentrated in 30 countries (see SM). India and China share the top two rankings of potential subsidence in terms of spatial extent and exposed population. Egypt and the Netherlands have the largest populations living in potential subsidence areas that are below the mean sea level. The greatest population densities in potential subsidence areas occur in Egypt and Indonesia, whereas the relative exposure per country, measured as the exposed population normalized by the total population, is greater than 30% for Egypt, Bangladesh, Netherlands, and Italy. The United States ranks first in terms of GDP exposed to potential subsidence, owing to its high GDP per capita.

Combination of the aforementioned metrics permits derivation of a potential subsidence index ranking (see SM). Seven of the first ten ranked countries have the greatest subsidence impact, accounting for the greatest amount of reported damages (Netherlands, China, USA, Japan, Indonesia, México and Italy).

During this century, climate change will cause serious impacts on the world's water resources through sea-level rise, more frequent and severe floods and droughts, changes in the mean value and mode of precipitation (rain versus snow), and increased evapotranspiration. Prolonged droughts will decrease groundwater recharge and increase groundwater depletion, intensifying subsidence. The global potential subsidence is predicted for 2040 using the same subsidence metrics and available global projections of water stress, water demand variations, climate, and population (see SM). Although predicted potential subsidence areas increase only by 7% globally, the threatened population is predicted to rise by 30%, affecting 1.6 billion inhabitants, 635 million of whom will be living in flood-prone ar-

eas. These changes will not be homogeneous. Between 2010 and 2040, the predicted population exposed to potential subsidence increases more than 80% in the Philippines, Iraq, Indonesia, México, Israel, Netherlands, Algeria, and Bangladesh. The increase will be moderate, less than 30%, for China, the United States, Italy, and Iran. Potential subsidence is forecasted to decrease in Japan and Germany, owing to effective groundwater management policies and population declines. Finally, potential subsidence is predicted to emerge in high-latitude northern countries like Canada and to increase in extent in Russia or Hungary, where climate change will favor longer dry seasons.

Further advancements in the global evaluation of subsidence can be made when a global historical database on subsidence rate, magnitude, and extent has been compiled, which could be largely sourced from continental monitoring of surface displacements using satellite radar imagery (8). Widespread continuous monitoring of subsidence will permit better evaluation of the potential impact of land subsidence, especially in countries like Indonesia, México, and Iran, where local studies revealed the highest subsidence rates worldwide, but the national dimension of subsidence is still unknown. Further research also is necessary to evaluate the cost of damage caused by current and historical subsidence worldwide. The combination of damage information with hazard estimates will permit improved assessments of potential loss and design of cost-effective countermeasures. Presently, annual subsidence costs are only published for China (US\$ 1.5 billion) and the Netherlands (US\$ 4.8 billion) (9). The greater subsidence costs in the Netherlands owe to the exposed population below the mean sea level and the large investments made to prevent flooding. Our model, which does not yet consider mitigation measures, likely overestimates potential subsidence exposure in the Netherlands and Japan, where groundwater management has effectively controlled subsidence over the past decades (10).

Our results identify 1596 major cities, or about 22% of the world's 7343 major cities that are in potential subsidence areas, with 57% of these cities also located in flood-prone areas. Moreover, subsidence threatens 15 of the 20 major coastal cities ranked with the highest flood risk worldwide (11), where potential subsidence can help delimit areas in which flooding risk could be increased and mitigation measures are necessary.

**“Subsidence threatens 15 of the 20 major coastal cities ranked with the highest flood risk worldwide.”**

Overall, potential global subsidence results can be useful to better define the spatial extent of poorly documented subsidence occurrences, discover unknown subsiding areas, prevent potential subsidence impacts wherever groundwater depletion occurs, and better identify areas where subsidence could increase the flooding risk. In any of these scenarios, an effective land-subsidence policy should include systematic monitoring and modeling of exposed areas, evaluation of potential damages, and cost-benefit analyses permitting implementation of adequate mitigation or adaptation measures. These measures should consider groundwater regulation and strategic long-term measures, such as the development of alternative water supplies and the protection and (or) enhancement of natural or artificial recharge of aquifers.

Considering that the potential subsidence may affect 635 million inhabitants living in flood-prone areas in 2040, it is of prime importance that potential subsidence is quantified and systematically included in flood risk analyses and related mitigation strategies. ■

#### REFERENCES AND NOTES

1. D. L. Galloway, T. J. Burbey, *Hydrogeol. J.* **19**, 1459 (2011).
2. J. S. Famiglietti, *Nat. Clim. Chang.* **4**, 945 (2014).
3. K. E. Trenberth, *Clim. Res.* **47**, 123 (2011).
4. J. P. M. Syvitski et al., *Nat. Geosci.* **2**, 681 (2009).
5. P. Döll, H. Müller Schmied, C. Schuh, F. T. Portmann, A. Eicker, *Water Resour. Res.* **50**, 5698 (2014).
6. R. G. Taylor et al., *Nat. Clim. Chang.* **3**, 322 (2013).
7. B. Jongman, P. J. Ward, J. C. J. H. Aerts, *Glob. Environ. Change* **22**, 823 (2012).
8. R. Lanari et al., *Remote Sens.* **12**, 2961 (2020).
9. T. H. M. Bucx, C. J. M. Van Ruiten, G. Erkens, G. De Lange in, *Proceedings of the International Association of Hydrological Sciences* **372**, 485 (2015).
10. K. A. B. Jago-on et al., *Sci. Total Environ.* **407**, 3089 (2009).
11. S. Hallegatte, C. Green, R. J. Nicholls, J. Corfee-Morlot, *Nat. Clim. Chang.* **3**, 802 (2013).
12. G. Herrera, P. Ezquerro, Global Subsidence Maps, figshare (2020); 10.6084/m9.figshare.13312070.

#### ACKNOWLEDGMENTS

Four anonymous peer reviewers and S. E. Ingebritsen (U.S. Geological Survey) helped to improve the manuscript. Funding for this study was provided partly by the Spanish Research Agency (AQUARISK, PRX19/00065, TEC2017-85244-C2-1-P projects) and PRIMA RESERVOIR project, and by all the institutions represented in the Land Subsidence International Initiative from UNESCO. G.H.-G., P.E., R.T., M.B.-P. and J.L.-V. designed the study, performed the analysis, and wrote the initial manuscript with input from all other authors. R.M.M., E.C.-C., and M.R. advised on the susceptibility analysis. R.M.M., J.L., P.T., and G.E. advised on hazard analysis. D.C.-F., J.L., P.T., E.C.C., G.E., D.G., W.C.H., N.K., M.S., L.T., H.W., and S.Y. advised on global exposure analysis. R.T., M.B.P., R.M.M., J.L., P.T., W.-C.H., N.K., L.T., H.W., and S.Y. contributed essential data for the analysis. All the authors edited and revised the manuscript through the different reviews. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. government. The authors declare no competing interests. All data included in this study are available at figshare (12).

#### SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/371/6524/34/suppl/DC1

10.1126/science.abb8549



BOOKS *et al.*

## PHILOSOPHY OF SCIENCE

# Science's irrational origins

Disputes in modern science are settled with empiricism alone, an approach early scholars would have questioned

By Itai Yanai<sup>1</sup> and Martin J. Lercher<sup>2</sup>

What is the scientific method, and what makes it the most efficient approach for generating insight? In *The Knowledge Machine*, Michael Strevens argues that to answer this question, we must acknowledge the role played by the undisciplined and emotional nature of the humans who carry it out. The book takes readers on a whirlwind tour through the history of science, rendering Arthur Eddington, Louis Pasteur, G. G. Simpson, Lord Kelvin, and many others as “warm-blooded organisms, whose enthusiasms, hopes, and fears mold their thinking far below the threshold of awareness.”

When asked what science is and how it functions, researchers offer a range of conflicting responses, notes Strevens. “Some scientists say that the essence of science is controlled or repeatable experiment, forgetting that experiments are of relatively little importance in cosmology or evolutionary biology. Some say advanced mathematical techniques are crucial, forgetting that the discoverers of genetics, for example, had no use for sophisticated math.”

The reviewers are at the <sup>1</sup>Institute for Computational Medicine, NYU Langone Health, New York, NY 10016, USA, and <sup>2</sup>Department of Biology and Institute for Computer Science, Heinrich Heine University, 40225 Düsseldorf, Germany. Email: itai.yanai@nyulangone.org

Strevens argues that an objective scientific method cannot exist, as all predictions from hypotheses rely on auxiliary assumptions such as the functioning of instruments, whose reliability must be evaluated subjectively. He proposes that the distinguishing feature of science is a procedural agreement, which he refers to as the “iron rule of explanation.” This rule holds that differences in scientific opinion must be settled by empirical testing alone. Thus, a scientist cannot argue for one hypothesis over another because it is more beautiful or more appealing philosophically or because it is better aligned with “God’s plan.” The iron rule applies only to official communications. Outside of such venues, scientists may think and believe as they wish.

That only data are capable of formally supporting a hypothesis may seem obvious, yet Strevens suggests that such an approach is inherently illogical. Imagine, for example, suggesting to Aristotle that he should restrict himself to data when arguing in favor of a particular theory. He would have pitied your ignorance. What better support for a theory could there be than an elegant chain of philosophical arguments?

Strevens argues that modern science owes its success to the relinquishing of deep philosophical understanding in favor of the shall-

Unlike the ill-fated bird rendered here, empirical inquiry thrives on deprivation, argues Strevens.

low power to predict empirical observations. As Isaac Newton—whom Strevens sees as the first truly modern scientist—wrote: “I have not as yet been able to deduce from phenomena the reason for these properties of gravity, and I do not feign hypotheses...It is enough that gravity really exists and acts according to the laws that we have set forth” (1).

Strevens proposes that scientists reason differently in public discourse and private venues. By drawing a clear distinction between formal scientific arguments and informal, behind-the-scenes scientific work, he provides a coherent framework for the divergent ideas of earlier philosophers of science: Karl Popper’s ideas on the falsification of hypotheses (2) form the basis of formal scientific discourse; Paul Feyerabend’s observations highlight the subjectivity of daily work, including the evaluation of assumptions (3); and the apparent security of a scientific paradigm guided by the iron rule compels scientists to perform elaborate experiments, thus generating data of otherwise unimaginable quantity and detail—a phenomenon described by Thomas Kuhn (4).

Strevens frames the toiling life of data generation as the cost scientists pay to gain access to the sacred halls of scientific excellence. What he overlooks is the supreme “pleasure of finding things out” (5). In his autobiography, French biologist François Jacob proposed the notion of “night science,” in which scientists generate new ideas and hypotheses in often unstructured thought processes (6). This approach, he argued, complements “day science,” wherein new ideas are tested empirically and reported formally. Thinkers such as Aristotle perceived day and night science as intertwined in a single process. Newton and his contemporaries founded modern science by separating them into distinct undertakings.

While Strevens’s iron rule may indeed be the foundation of modern science’s success, the methods scientists use to come up with new ideas remain elusive. ■



**The Knowledge Machine**  
Michael Strevens  
Liveright, 2020. 368 pp.

## REFERENCES AND NOTES

1. I. Newton, *The Mathematical Principles of Natural Philosophy* (Benjamin Motte, 1687).
2. K. Popper, *The Logic of Scientific Discovery* (Hutchinson, 1959).
3. P. Feyerabend, *Against Method* (Verso Books, 1975).
4. T. Kuhn, *The Structure of Scientific Revolutions* (Univ. of Chicago Press, 1962).
5. R. Feynman, *The Pleasure of Finding Things Out* (Perseus Books, 1999).
6. F. Jacob, *The Statue Within* (Cold Spring Harbor Laboratory Press, 1995).

10.1126/science.abf4887



## PHYSICS

# Thermodynamics and the matter of life

A scientist considers life's genesis through the physics of Exodus

By Eran A. Mukamel<sup>1</sup> and Amelia M. Glaser<sup>2</sup>

**W**hat is life? Seventy-five years after Erwin Schrödinger took up this fundamental human question (1), another physicist, Jeremy England, offers a bold update. His new book, *Every Life Is on Fire*, explores the physics of what makes some configurations of matter lifelike and others inert.

England writes sensitively about biological complexity from the molecular to the human scale. He also goes a step further, connecting the thermodynamics of life with meditations on its moral and ethical implications through the lens of the Bible. England is convinced that scientific and humanistic outlooks can inform and enrich each other rather than locking horns in epistemological opposition. If some scientists (not to mention Bible interpreters) find the citation of sacred texts in a scientific work heretical, the book's lucid explanations of nonequilibrium thermodynamics and biophysics will nevertheless prompt the old question: What about living matter inspires wonder and a sense of greater significance?

Schrödinger famously anticipated the discovery of DNA's structure, arguing on physical principles that the stability of inherited traits in living organisms requires the existence of an "aperiodic crystal" as the genetic material. While he accepts the importance of molecular genetics and Darwinian evolution, England is more interested in the macroscopic, plainly visible hallmarks of life: self-replication, ordered and functionally specialized structures, and the ability to harness energy sources and to predict and respond to the environment. He argues that each of these capacities can be understood through physical processes that occur in ordinary, lifeless materials when driven by an energy source.

England explains the physics through accessible analogies. Schrödinger's concept

of a crystal evokes stable equilibrium states akin to the surface of a frozen lake. England shows how nonequilibrium systems, like sand dunes piled up by a desert wind or a river network carved by the cyclical flow of water, are better physical analogs for adaptable, robust, and self-organizing living structures.

The book builds up to a discussion of recent theoretical work by England and his collaborators, which argues that matter—living or not—will, in some circumstances, evolve over time to efficiently harness specific environmental energy sources. England's theory, called dissipative adaptation,



Moses' staff-turned-serpent illustrates the line between living and nonliving.

does not seek to replace Darwinian evolution. It does, however, propose that at least some distinctive features of life could arise through a nonbiological process of selection. According to this theory, chemical or physical structures that efficiently use a source of energy to reduce disorder without themselves being destroyed are more likely to catalyze the formation and growth of similar structures.

This argument—the most speculative part of the book—leaves open key questions, such as how selection can proceed without a biological mechanism to faithfully encode and reproduce specific structures. It would have been stronger had England included concrete experimental

**Every Life Is on Fire:  
How Thermodynamics  
Explains the Origins  
of Living Things**

Jeremy England  
Basic Books, 2020. 272 pp.



predictions that could falsify dissipative adaptation as a driver of biological self-organization. Still, his framing of non-Darwinian evolutionary processes with the language of nonequilibrium thermodynamics reveals how some remarkable features of biological order can be found in

nonliving physical systems.

The distinction between living and nonliving matter is not an outcome of scientific inquiry, it is a conceptual framework that humans bring to our encounters with the world. As such, a preoccupation with the boundaries that define life is shared not only by physicists and biologists but also by those concerned with cultural traditions and religious practices. England uses images of life's boundaries from the biblical Book of Exodus, where staffs turn into serpents and a bush burns without being consumed, to evoke the fuzzy line between living and nonliving matter. Such references help explain, in memorable and human terms, the physics of this elusive boundary.

It is rare for modern science to engage ancient religious texts;

these traditions are more often nonoverlapping magisteria (2), if not fundamentally incompatible. *Every Life Is on Fire* shows that scripture can enrich our scientific interest in living systems, providing an ethical, moral, and even spiritual context. For the reader willing to brave metaphorical land mines, there is much to be learned by exploring the border regions, whether between physics and biology, between science and religion, or between life and lifeless matter. ■

## REFERENCES AND NOTES

1. E. Schrödinger, *What Is Life? The Physical Aspect of the Living Cell* (The University Press, 1945).
2. S. J. Gould, *Nat. Hist.* **106**, 16 (1997).

10.1126/science.abf5633

The reviewers are at the <sup>1</sup>Department of Cognitive Science and <sup>2</sup>Department of Literature, University of California, San Diego, La Jolla, CA 92093, USA. Email: emukamel@ucsd.edu

# RESEARCH

## IN SCIENCE JOURNALS

Edited by **Michael Funk**

### CLOUD PHYSICS

#### Aerosols give clouds a lift

It has been observed that atmospheric aerosols can strengthen updrafts in deep convective clouds such as those that form in thunderstorms. Past work has linked such invigoration with the latent heat released by water condensation or freezing in chains of processes that depend on aerosol concentrations. Abbott and Cronin suggest a third possibility in which updrafts intensify because high aerosol concentrations increase environmental humidity by mixing more condensed water into the surrounding air, which in turn favors stronger updrafts. —HJS *Science*, this issue p. 83

Atmospheric aerosols invigorate thunderstorms by increasing humidity and thus stimulating convection.

### CELL CYCLE

#### Checking fidelity in cell division

Everything has to go right during cell division, so a checkpoint mechanism known as the spindle-assembly checkpoint prevents mitosis from proceeding unless the kinetochores that attach chromosomes to the spindle microtubules are properly engaged. Two papers now reveal the detailed molecular choreography that allows a single, unattached kinetochore to arrest cell division: Lara-Gonzalez *et al.* used a visual probe that tracks a specific form of one of the checkpoint complex proteins, and Piano *et al.* used a biochemical reconstitution of the checkpoint. Together, these studies reveal how protein interaction, spatial constraints, phosphorylation,

and catalytic conversion of the protein Mad2 to its active form allow this all-important sensor to function. —LBR

*Science*, this issue p. 64, p. 67

### PROTEIN FOLDING

#### One sequence encoding two structures

Most proteins have stable, folded structures, but there are rare examples of metamorphic proteins that can switch between two different folds that may each have a different function. Dishman *et al.* investigated the evolution of XCL1, which is a member of the chemokine family that interconverts between the chemokine fold and a second, noncanonical fold that forms dimers. The authors used nuclear magnetic resonance

spectroscopy to investigate the structures of inferred evolutionary ancestral sequences. Their results suggest that XCL1 evolved from an ancestor with the chemokine fold and then transitioned to prefer the non-canonical fold before reaching the modern-day metamorphic protein. —VV

*Science*, this issue p. 86

### MATERIALS SCIENCE

#### Stretching diamond to the limit

Diamond is thought of as being unbendable, but thin samples can actually deform elastically. Applying relatively large amounts of strain to diamond may shift its electronic properties, which is of interest for a number of applications. Dang *et al.* elastically stretched

micrometer-sized plates of diamond along different crystallographic directions. These relatively large samples show that deep-strain engineering can be accomplished in more uniform diamond specimens and may have a large impact on the electronic properties. —BG

*Science*, this issue p. 76

### PROTEIN SYNTHESIS

#### Co-co assembly for oligomers

Most of the human proteome forms oligomeric protein complexes, but how they assemble is poorly understood. Bertolini *et al.* used a ribosome-profiling approach to explore the existence of a cotranslational assembly mode based on the interaction of two nascent polypeptides, which they



call the “co-co” assembly. Proteome-wide data were used to show whether, when, and how efficiently nascent complex subunits interact. The findings also show that human cells use co-co assembly to produce hundreds of different homo-oligomers. Co-co assembly involving ribosomes translating one messenger RNA may resolve the long-standing question of how cells prevent unwanted interactions between different protein isoforms to efficiently produce functional homo-oligomers. —SMH

*Science*, this issue p. 57

## ACTIVE MATTER

### Shake, rattle, and help each other along

In classical statistical mechanics, the deterministic dynamics of a many-body system are replaced by a probabilistic description. Chvykov *et al.* work toward a similar description for the nonequilibrium self-organization of collectives of active particles. In these systems, continuously input energy drives localized fluctuations, but larger-scale ordering can emerge, such as in the flight of a flock of birds. A key concept in their theory is the importance of rattling, whereby ordered patterns emerge through local collisions between neighbors at specific frequencies. The authors demonstrate this behavior using



A trio of robots capable of emulating collective behaviors

a set of flapping robots and produce related simulations of the robot behavior. —MSL

*Science*, this issue p. 90

## TUMOR IMMUNOLOGY

### Autophagy protects tumors from T cells

Tumors evade antitumor T cells by various mechanisms. Young *et al.* used a CRISPR screen to show that tumor necrosis factor- $\alpha$  (TNF $\alpha$ ) and autophagy play a role in the T cell-mediated killing of tumor cells. Pharmacologic or genetic inhibition of autophagy in tumor cells increased TNF $\alpha$ -mediated T cell killing of tumor cells. Deletion of the gene *Rb1cc1* in tumor cells improved the efficacy of immune checkpoint blockade in a mouse tumor model. However, deleting the TNF $\alpha$  receptor in tumor cells partially abrogated the improved efficacy of immune checkpoint blockade in the absence of *Rb1cc1*. Thus, autophagy inhibition may improve T cell-mediated immunotherapies in patients who have cancer. —DAE

*Sci. Immunol.* **5**, eabb9561 (2020).

## NEURODEGENERATION

### Saving neurons in Parkinson's disease

Parkinson's disease is characterized by the progressive loss of dopaminergic neurons that leads to loss of motor control and cognitive decline. Kim *et al.* found that activating the kinase Akt1, such as with chlorogenic acid (a polyphenol found in coffee), prevented both neuronal death and motor and cognitive impairments in two mouse models of Parkinson's disease. Akt1 inhibited neuronal death by transcriptionally activating a gene involved in regulating programmed cell death. The clinical relevance of this mechanism was supported by correlative data collected from postmortem patient brain tissue. —LKF

*Sci. Signal.* **13**, eaax7119 (2020).

## IN OTHER JOURNALS

Edited by **Caroline Ash** and **Jesse Smith**



## NEUROSCIENCE

### Oscillations around memory

Hippocampal oscillations in the theta range have been hypothesized to play a central role in organizing neuronal ensembles to link together item and contextual representations. Experimental evidence in rodents shows the importance of theta oscillations for associative memory. However, the role of hippocampal theta oscillations in human memory is not as well understood. Kota *et al.* administered an associative recognition memory task to epilepsy patients who happened to have electroencephalogram electrodes implanted for other medical reasons. Theta oscillatory power increase in the 2- to 5-Hz range and phase reset in the hippocampus reflected processes supporting recollection, rather than familiarity, during encoding and retrieval. These observations link theta-range activity to associative

memory encoding and retrieval in humans. —PRS

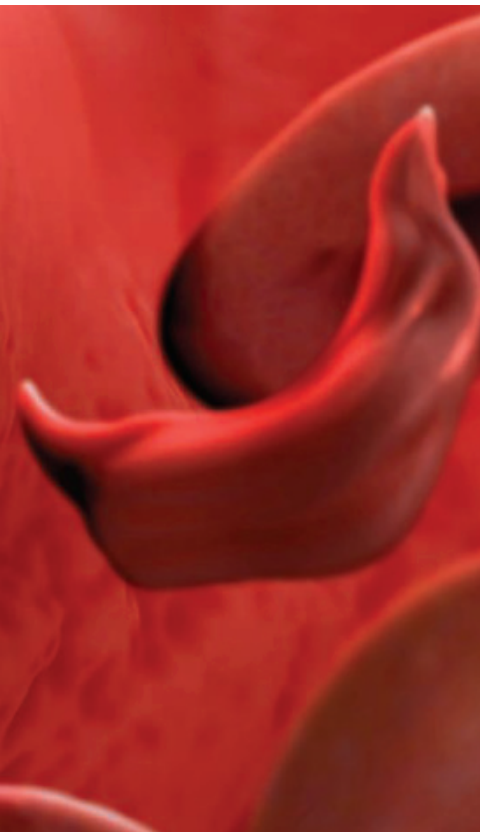
*J. Neurosci.* **40**, 9507 (2020).

## TYPE 1 DIABETES

### Interfering with diabetes

Insulin, discovered a century ago, remains the mainstay of treatment for type 1 diabetes, as autoimmunity destroys the patients' insulin-producing pancreatic  $\beta$  cells. A study by Quattrin *et al.* indicates that it may be possible to slow down the progression of this disease. In this phase 2 clinical trial, young patients with newly diagnosed type 1 diabetes were treated with golimumab, an antibody against the cytokine tumor necrosis factor- $\alpha$ , and compared with a placebo group. Although golimumab did not fully prevent disease progression, it produced partial remissions and decreased the amount of insulin required, offering a potentially promising addition to the therapeutic options for this patient group. —YN

*N. Engl. J. Med.* **383**, 2007 (2020).



## GENE THERAPY

### Editing blood disorders

**T**he two most common monogenic diseases, transfusion-dependent  $\beta$ -thalassemia (TDT) and sickle cell disease (SCD), result from mutations in the hemoglobin  $\beta$ -subunit gene (*HBB*), which is an essential element of adult hemoglobin A ( $\alpha_2\beta_2$ ).

Current therapies for TDT and SCD are limited and do not address their underlying causes. Frangoul *et al.* report the treatment of two patients (one with TDT and the other with SCD) using gene therapy. After myeloablation, the patients were infused with their own hematopoietic stem and progenitor cells subjected to CRISPR-Cas9 gene editing of the erythroid-specific enhancer region of BCL11A. This transcription factor represses the expression of  $\gamma$ -globin, a component of  $\alpha_2\gamma_2$  fetal hemoglobin that is known to ameliorate the severity of these disorders. More than a year later, both patients showed sustained engraftment of edited cells in the blood and bone marrow and increased fetal hemoglobin expression, which relieved symptoms and obviated the need for transfusions.

—STS *N. Engl. J. Med.* **10.1056/NEJMoa2031054** (2020).

Red blood cells (shown here in an artist's rendering) become distorted in sickle cell disease because of abnormal hemoglobin, but the symptoms can be eased by gene editing.

the search for alternatives to synthetic, nondegradable polymers. Bagasse is the dry fibrous residue from sugarcane processing that is primarily used as a fuel source because its fibers are short and thus mechanically weak. Liu *et al.* developed a fiber-hybridization process to combine bagasse with long bamboo fibers to form a tighter intertwined network. Alkyl ketene dimer was added to modify the cellulose fibers to make the blended material more resistant to oil and water. The blend can be shaped into cups and food containers suitable for holding beverages or food for short periods, and the packaging will notably degrade within 60 days of burial in normal soil. —MSL

*Matter* **3**, P2066 (2020).

## ARCTIC SEA ICE

### A closer look at loss

As the climate has warmed, most markedly in high northern latitudes, the amount of Arctic sea ice remaining at the end of the summer melting season has decreased in response. To project a more detailed picture of sea ice loss over the remainder of the century, Arthun *et al.* conducted model simulations of seasonal and regional variations under different climate paths. They predict that sea ice loss will take place in all regions and all seasons; that all Arctic shelf seas will become ice free in the summertime by the mid-2050s, even in a low-emission scenario; and that parts of the Arctic may lose even their winter sea ice before the year 2100. —HJS

*Geophys. Res. Lett.*

**10.1029/2020GL090825** (2020).

## SELF-ASSEMBLY

### Ring in 23 building blocks

Bottom-up synthesis of specific oligomers often relies on steadily increasing symmetry. Pairwise interactions between building blocks repeat themselves throughout the structure. Pappas *et al.* report a case of self-assembly that breaks this paradigm. As many as 23 monomers selectively combined to form low-symmetry macrocycles stabilized by a self-coiled geometry. The building blocks consist of dimercaptobenzene units that can reversibly link up through disulfide bonds, each bearing a dipeptide tail that guides the assembly process through hydrogen bonding and hydrophobic interactions. —JSY

*Nat. Chem.* **12**, 1180 (2020).

## CELL BIOLOGY

### Here comes the cavalry!

Healing small scratch wounds in the skin is much less well understood than the closing of

deeper wounds. Bornes *et al.* used intravital microscopy to study the healing of scratches on the backs of mice. The authors followed individual keratinocytes as the cells multiplied and migrated to rapidly reestablish the skin's protective barrier. Basal keratinocytes multiplied at sites somewhat distant from the wound and then swarmed toward the wound, moving as a sheet of independently migrating cells. This allowed the replacement cells to bypass any intervening obstacles such as intact immobile hair follicles or sweat glands. The findings challenge a model proposing that the

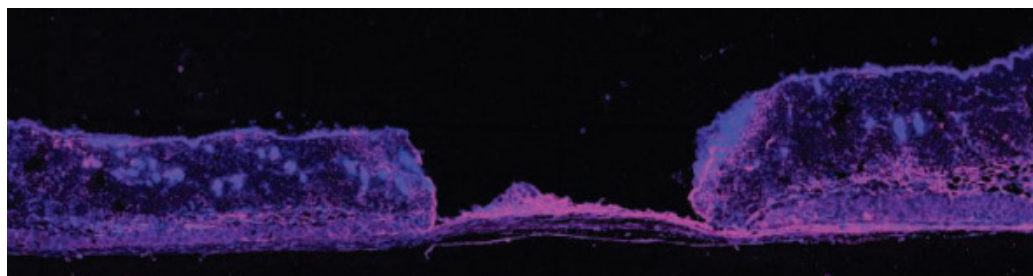
leading edge of migrating cells directly transmits forces to adjacent keratinocytes more distant from the wound bed to physically drag them along. Instead, leading-edge keratinocytes may initiate wound closure followed by individually moving keratinocytes within a cohesive layer. —SMH

*Life Sci. Alliance* **4**, e202000765 (2020).

## MATERIALS SCIENCE

### Food packaging from sugarcane and bamboo

The push to reduce disposable materials has encouraged



Immunofluorescent image of mouse skin 8 hours after wounding, showing sheets of migrating keratinocytes plugging the site



## REVIEW SUMMARY

## CIRCADIAN RHYTHMS

## Clocks, cancer, and chronochemotherapy

Aziz Sancar\* and Russell N. Van Gelder\*

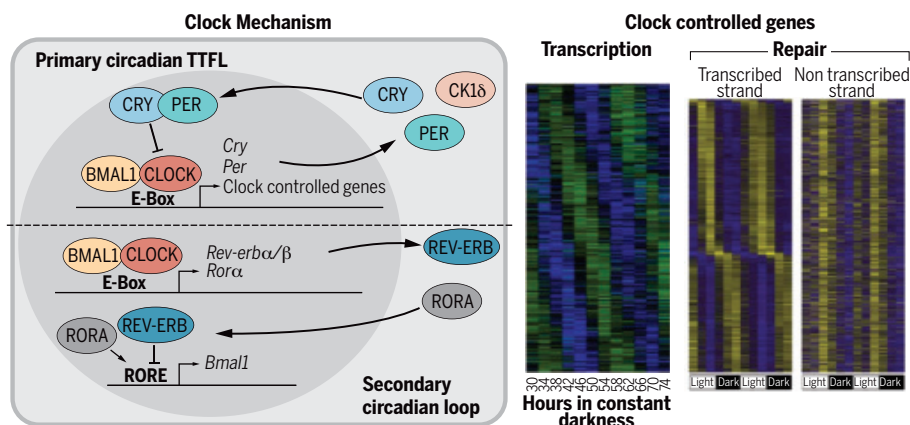
**BACKGROUND:** The core of the mammalian circadian clock mechanism is a time-delayed transcription-translation feedback loop (TTFL), which influences the transcription and expression of a large fraction of the transcriptome. Through this mechanism, the mammalian circadian clock modulates many physiological functions, including the timing of cell division and rates of metabolism in specific tissues. Circadian clock dysfunction is associated with several human disease states, including jet lag and sleep phase disorders, and it likely contributes significantly to the development of metabolic syndrome. With respect to cancer, animal studies have suggested that specific carcinogenic mechanisms, such as ultraviolet radiation for skin cancer, have a strong circadian rhythm. Epidemiologic studies have yielded conflicting results as to whether circadian clock disruption by night or shift work is carcinogenic. In animal studies, tumors grafted into

animals with disrupted rhythms grow more rapidly than those grafted into control animals. Studies of mice genetically lacking specific components of the circadian clock show increased rates of tumorigenesis for certain clock genes and certain tumors but show reduced rates for other clock genes. Similarly, the response to chemotherapy may also vary with time of day, which has led to enthusiasm for chronochemotherapy as a means to improve the therapeutic efficacy of cancer treatment while limiting toxicity. However, clinical trials of chronochemotherapy have generally not shown improved efficacy and have even shown worse outcomes in subsets of patients compared with conventionally timed therapies.

**ADVANCES:** Polymorphisms in circadian clock genes including *Npas2* and *Clock* have been identified in genome-wide association studies as relatively weak but significant modifiers of

breast cancer incidence, and core circadian clock gene expression is frequently dysregulated in human tumors. However, it is not possible to generalize that loss of the clock leads to increased cancer incidence, as some clockless animals actually show resistance to specific cancer pathways (e.g., *Cryptochrome*-less mice are resistant to p53 mutation-induced tumors). In other cases, different clock gene mutations result in opposite phenotypes with respect to carcinogenesis for the same tumor type. Perhaps the best-studied mechanistic interaction between circadian clock and carcinogenesis involves studies of the circadian rhythms of nucleotide excision DNA repair. Although basal excision repair has a circadian rhythm with a specific maximal phase, the rhythm of an individual gene's repair is dependent on the phase of that gene's transcriptional rhythm; there is no single phase at which DNA is generally more or less easily repaired. Other notable advances in the field include the demonstration of direct mechanistic linkage of c-MYC expression to circadian clock control and the demonstration that oncogenes *c-Myc*, *p53*, and *Ras* all affect the circadian core TTFL, consistent with the finding that the circadian clock of tumors is frequently dysregulated.

**OUTLOOK:** Tumorigenesis is clearly affected by circadian mechanisms, but the hypothesis that circadian clock genes are general tumor suppressors is not supported. Rather, specific tumors and their underlying mechanisms are differentially affected by the function of specific clock genes. Conversely, specific oncogenes may cause dysregulation of the circadian clock in tumors; the pathogenic significance of the dysregulated clock in tumors is not fully understood. The example of circadian control of DNA nucleotide excision repair illuminates the challenges in exploiting the interaction between clocks and cancer clinically, as the phase of circadian susceptibility to DNA damage varies for each gene on the basis of its underlying transcriptional rhythm. Although the concept of chronochemotherapy is attractive, the complexities of clock-cancer interactions make prediction of the effects of timed drug administration challenging. Mistiming of chemotherapeutic agents has the potential to be harmful. As chemotherapeutic agents increase in specificity, the circadian effects of administration may be better understood and optimized by understanding the specific interactions between the circadian clock mechanism and therapeutic targets. ■



**Mammalian circadian clock controls transcription and DNA repair.** (Left) The mammalian circadian clock mechanism is a time-delayed TTFL. BMAL1-CLOCK constitute the positive arm and cryptochrome (CRY1 and CRY2)–period (PER1 and PER2) constitute the repressive arm; the primary feedback loop is consolidated by a secondary loop made up of REV-ERB $\alpha$  inhibitor and ROR $\alpha$  activator. In a given tissue, ~10% of the genes are expressed with significant circadian (near–24 hour) periodicity. (Right) Effect of the clock on transcription and nucleotide excision repair in mice is shown in two heatmaps, where the green (transcription) and yellow (repair) represent the intensity of the signal. The left-side expression heatmap shows 854 clock-controlled transcripts in the livers of mice kept in the dark for 44 hours [adapted from B. H. Miller *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 3342 (2007). Copyright (2007) National Academy of Sciences]. Each of these genes is expressed maximally at a specific time of day. The right-side repair heatmap shows 1661 genes from the kidneys of mice kept under a 12-hour light–12-hour dark condition treated with the chemotherapeutic cisplatin [adapted from Y. Yang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **115**, E4777 (2018), Copyright (2018) National Academy of Sciences]. Damage is quantified for both transcribed strands (TS) and nontranscribed strands (NTS). The NTS shows a monophasic rhythm for all genes with a peak in early evening. For the TS, each gene shows a specific maximum for repair during the cycle corresponding to its peak phase of transcription. The complexity of individual gene repair timing creates substantial challenges for optimizing circadian timing of chemotherapy administration.

The list of author affiliations is available in the full article online.

\*Corresponding author. Email: aziz\_sancar@med.unc.edu (A.S.); russvg@uw.edu (R.N.V.G.)

Cite this article as A. Sancar, R. N. Van Gelder, *Science* **371**, eabb0738 (2021). DOI: 10.1126/science.abb0738

**S** READ THE FULL ARTICLE AT  
https://doi.org/10.1126/science.abb0738

## RESEARCH ARTICLE SUMMARY

## STRUCTURAL BIOLOGY

## Structural basis of antagonizing the vitamin K catalytic cycle for anticoagulation

Shixuan Liu, Shuang Li, Guomin Shen, Narayanasami Sukumar, Andrzej M. Krezel, Weikai Li\*

**INTRODUCTION:** Vitamin K antagonists (VKAs), such as warfarin, are oral anticoagulants commonly used to treat and prevent thromboembolic diseases, including stroke and heart attack. Vitamin K supplementation, on the other hand, has saved the lives of numerous newborns with deficient hemostasis. Central to these therapeutic processes is vitamin K epoxide reductase (VKOR), an endoplasmic membrane enzyme that generates the active form of vitamin K to support blood coagulation. VKAs inhibit VKOR catalysis, which is carried

out by two cysteine pairs, one directly reducing substrates and another mediating electron transfers. VKOR homologs and paralogs (VKOR-like) constitute a large family of integral membrane thiol oxidoreductases, but understanding their catalytic and inhibitory mechanisms has been challenging in the absence of high-resolution structures.

**RATIONALE:** Overdose of VKAs often causes major or fatal bleeding, accounting for one-third of hospitalizations for all adverse drug

reactions in older adults. Despite decades of clinical difficulties, mechanistic insights are lacking for the antagonism of VKAs and for their target enzyme VKOR, which manages to transfer electrons across the water-membrane interface to support its distinct activity of epoxide reduction. Here, we present 11 crystal structures of human VKOR and a VKOR-like paralog with various substrates and antagonists in different functional states, revealing nearly the entire catalytic cycle of VKOR enzymes and the action of VKAs.

**RESULTS:** Structures with four representative VKAs show that their hydrogen bonding with Asn<sup>80</sup> and Tyr<sup>139</sup> provides the recognition specificity in a largely hydrophobic pocket. This VKA-binding pocket, also serving as the active site, is surrounded by a four-transmembrane-helix bundle and covered by a cap domain. Mutations that destabilize the cap domain or directly disrupt the warfarin-binding interactions result in warfarin resistance. Metabolic inactivation of warfarin is through a hydroxyl modification that is energetically unfavorable in the hydrophobic pocket. High-potency inhibition by “superwarfarins” is afforded by their large side groups that bind to a tunnel designated for the isoprenyl chain of vitamin K. Structural comparison with a ligand-free state suggests that local binding interactions of warfarin lead to a global change from open to closed protein conformation.

Substrate-bound structures reveal that a reduced active-site cysteine is required to form a charge-transfer complex or covalent complex. These substrate adducts form hydrogen bonds with Asn<sup>80</sup> and Tyr<sup>139</sup> to facilitate the catalytic chemistry, and their stable binding interactions induce the closed protein conformation. This conformational change brings together the two pairs of cysteines, triggering electron transfer that enables the reduction of substrates.

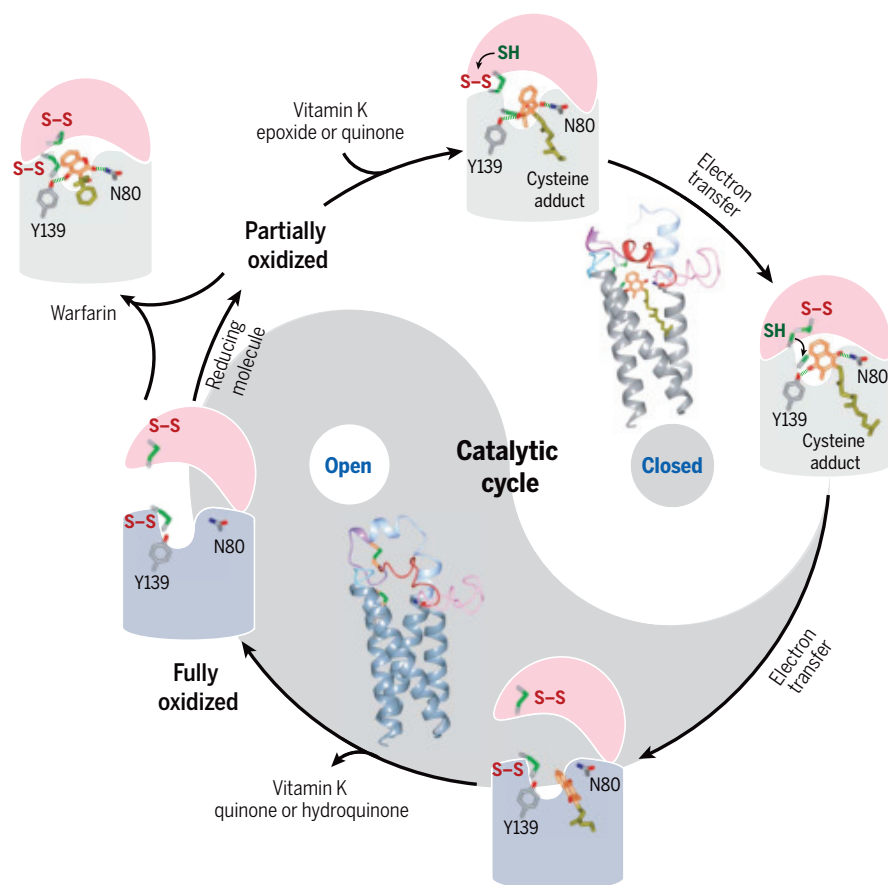
**CONCLUSION:** The structures reveal an activation mechanism in which stably bound substrate adducts trigger restructuring of VKOR to promote electron transfer. The high potency of VKAs results from mimicking the conformational change at the global level and the substrate hydrogen-bonding interactions at the local level. These mechanistic insights provide the basis to design new therapeutic strategies of anticoagulation. ■

The list of affiliations is available in the full article online.

\*Corresponding author. Email: weikai@wustl.edu

Cite this article as S. Liu *et al.*, *Science* **371**, eabc5667 (2021). DOI: 10.1126/science.abc5667

**S** READ THE FULL ARTICLE AT  
https://doi.org/10.1126/science.abc5667



**The catalytic cycle and inhibition of VKOR.** Partially oxidized VKOR forms a cysteine adduct with substrates, vitamin K epoxide, or quinone, whose binding induces a closed conformation, juxtaposing all cysteines (S-S or SH) for unimpeded electron transfer. VKOR becomes fully oxidized with an open conformation that releases reaction products, vitamin K quinone, or hydroquinone. Warfarin locks VKOR in both redox states into the closed conformation. The luminal and transmembrane domains of VKOR are shown as a pink hemisphere and gray cylinder, respectively. Y139, Tyr<sup>139</sup>, N80, Asn<sup>80</sup>.



## RESEARCH ARTICLE SUMMARY

## TRANSCRIPTION

Steps toward translocation-independent RNA polymerase inactivation by terminator ATPase  $\rho$ 

Nelly Said\*, Tarek Hilal\*, Nicholas D. Sunday, Ajay Khatri, Jörg Bürger, Thorsten Mielke, Georgiy A. Belogurov, Bernhard Loll, Ranjan Sen, Irina Artsimovitch†, Markus C. Wahl†

**INTRODUCTION:** Factor-dependent transcription termination is essential to limit pervasive transcription, maintain genome stability, balance the expression of neighboring genes, and recycle RNA polymerase (RNAP). Two main classes of models can explain how termination factors stop RNA synthesis. In RNA-centric models, a terminator, powered by adenosine triphosphate (ATP)-dependent RNA translocase activity or by exonucleolytic RNA degradation, moves along the nascent RNA and rear-ends RNAP, dislodging it from the RNA. In transcription elongation complex (EC)-centric models, a terminator induces conformational changes in RNAP that inactivate it. Evidence in support of both mechanisms exists for translocases and exonucleases that elicit termination in bacteria and eukaryotes, but molecular details of their actions remain elusive because, once committed to termi-

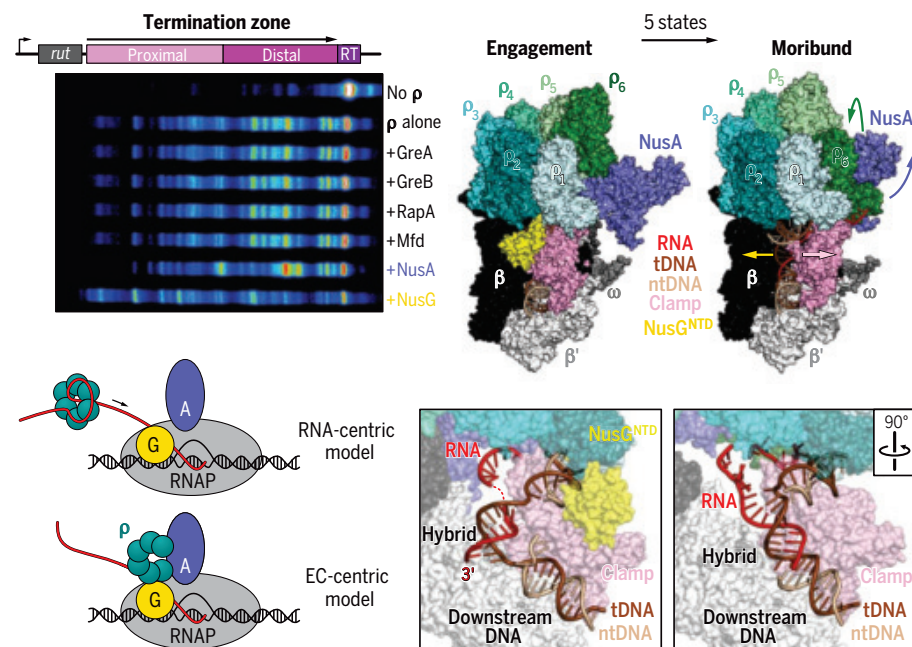
nation, transcription complexes disassemble rapidly and are thus refractory to structure/function analyses.

**RATIONALE:** To elucidate the structural basis for termination, we used the archetypal ring-shaped hexameric helicase  $\rho$ . *Escherichia coli*  $\rho$ , perhaps the strongest molecular motor known, can load onto free RNA as an open ring, close the ring around the RNA, and engage in ATP-dependent translocation, removing any obstacle from its path. During transcription,  $\rho$  triggers RNA release from the EC within a well-defined termination zone once ~90 nucleotides of C-rich RNA, which  $\rho$  binds with high affinity, have been synthesized by RNAP. We surmised that a  $\rho$ -bound EC poised to enter this termination zone will be metastable, giving rise to an ensemble of intermediates en route to termination. We used single-particle cryo-

electron microscopy (cryo-EM) to analyze these “peri-termination” *E. coli* ECs bound to  $\rho$ , an ATP analog, and general elongation factors NusA and NusG known to modulate  $\rho$  activity. We also carried out in vitro and in vivo functional assays to validate key interactions suggested by our structural analysis.

**RESULTS:** We report the structures of seven intermediates along the termination pathway.  $\rho$  is recruited to the EC via extensive contacts to RNAP, NusA, and NusG, but initially makes no contacts with RNA. After recruitment, rearrangements of the  $\rho$  hexamer, NusA, upstream DNA, and several regions of RNAP set up a stage for RNA engagement by  $\rho$ . The N-terminal zinc-binding domain of the RNAP  $\beta'$  subunit aids  $\rho$  in capturing the nascent RNA, a synergy that is supported by in vivo analysis of  $\rho$  and  $\beta'$  mutants. Upon anchoring the RNA,  $\rho$  induces structural rearrangements that lead to the displacement of NusG and weakening of the RNAP grip on nucleic acids due to partial opening of the  $\beta'$  clamp domain. The formation of a moribund complex, in which the clamp is wide open and the RNA is dislodged from the active site, completes the RNAP inactivation by  $\rho$ . Remarkably, the  $\rho$  ring is held open by the network of  $\rho$  contacts with RNAP and NusA throughout the entire pathway, preventing  $\rho$  from exerting force on RNA. Our data argue that  $\rho$  travels with RNAP rather than chases after it, and that termination is favored by pause-promoting conformational changes in the EC rather than by the reduced rate of RNA synthesis.

**CONCLUSION:** This study explains how  $\rho$  is targeted to RNAs that are still being made and cooperates with NusA and NusG to effect striking conformational changes that inactivate the transcribing RNAP. Hitchhiking on RNAP enables  $\rho$  to survey and silence useless and harmful transcripts independently of their sequence, as documented for several bacterial  $\rho$  orthologs. Unexpectedly,  $\rho$  stalls transcription without engaging its powerful motor activity, which may be essential after termination to destroy R-loops, the toxic by-products of the EC dissociation. A growing list of allosteric mechanisms of transcription regulation suggests that many accessory factors may exploit dynamic properties of RNAP to modulate RNA synthesis, acting together with the orthologs/analogs of Nus factors present in all domains of life. ■



**$\rho$  traps NusA/NusG-modified elongation complexes in a moribund state.** NusA and NusG are the only general transcription factors in *E. coli* that modulate  $\rho$ -dependent termination. Conflicting models explain how  $\rho$  terminates RNA synthesis. Cryo-EM analysis of  $\rho$ /NusA/NusG-ECs and structure-informed biochemical analyses support an EC-centric model, revealing how an initial engagement complex is converted stepwise to a moribund complex. The pathway involves rearrangements of  $\rho$ , NusA/G, and RNAP elements, and culminates in a massive displacement of the RNA 3'-end from the RNAP active site.

The list of author affiliations is available in the full article online.  
\*These authors contributed equally to this work.

†Corresponding author. Email: artsimovitch.1@osu.edu (I.A.); markus.wahl@fu-berlin.de (M.C.W.)  
Cite this article as N. Said et al., *Science* 371, eabd1673 (2021). DOI: 10.1126/science.abd1673

**S** READ THE FULL ARTICLE AT  
<https://doi.org/10.1126/science.abd1673>

## RESEARCH ARTICLE SUMMARY

## STRESS RESPONSES

# QRICH1 dictates the outcome of ER stress through transcriptional control of proteostasis

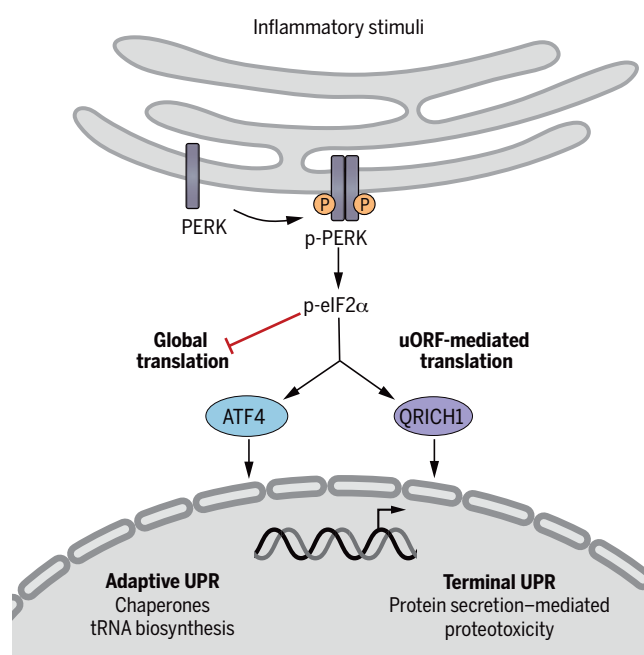
Kwontae You, Lingfei Wang, Chih-Hung Chou, Kai Liu, Toru Nakata, Alok Jaiswal, Junmei Yao, Ariel Lefkovich, Abdifatah Omar, Jacqueline G. Perrigoue, Jennifer E. Towne, Aviv Regev\*, Daniel B. Graham\*, Ramnik J. Xavier\*

**INTRODUCTION:** Tissue homeostasis requires the coordinated activity of multiple cell types to initiate and then resolve inflammation. Endoplasmic reticulum (ER) stress is a hallmark of inflammation and exacerbates tissue pathology across a broad range of human diseases. Environmental stressors associated with inflammation and cell-intrinsic metabolic demands can elicit ER stress, protein misfolding, and cell death. To counteract these processes, stress response pathways, including the unfolded protein response (UPR), facilitate adaptation to stress and tissue restitution. Cells sense ER stress and initiate the UPR through three coordinated pathways mediated by the effector proteins inositol-requiring enzyme 1 (IRE1/ERN1), activating transcription factor 6 (ATF6), and protein kinase RNA-like ER kinase (PERK, EIF2AK3). Collectively, the UPR effector pathways fine-tune the rate of protein translation and induce transcriptional up-regulation of genes that promote ER function, such as those encoding chaperone proteins and secretory machinery. Although these functional responses to ER stress by the UPR pathways aim to restore cellular homeostasis, prolonged and unresolved ER stress can elicit programmed cell death. In this context, the molecular mechanisms that dictate the outcome of ER stress are incompletely understood.

**RATIONALE:** Mismanagement of ER stress in intestinal epithelial cells can lead to disruption of barrier integrity, resulting in exposure of the host immune system to commensal microbes that trigger uncontrolled inflammation. With accumulating evidence highlighting the prominent role of ER stress in disease, it remains to be determined how the UPR directs divergent cell fate decisions. The UPR either induces an adaptive phase that promotes recovery of ER proteostasis and cell survival or induces a

terminal phase that initiates the active engagement of programmed cell death pathways.

**RESULTS:** Toward the objective of defining mechanisms controlling the adaptive versus terminal UPR, we used single-cell RNA sequencing (scRNA-



**QRICH1 controls a distinct arm of the PERK-eIF2 axis to modulate proteostasis and dictate entry into the adaptive versus terminal UPR.** In response to ER stress, PERK phosphorylates eIF2α, suppressing global translation while promoting ATF4 and QRICH1 translation by bypassing inhibitory upstream open reading frames (uORFs). QRICH1 localizes to the nucleus and positively regulates the transcription of genes that regulate protein secretion. Prolonged QRICH1 expression is associated with proteotoxicity and cell death during the terminal UPR, whereas its down-regulation is associated with restoration of ER homeostasis during the adaptive UPR. P, phosphorylation; p-PERK, phosphorylated PERK; p-eIF2α, phosphorylated eIF2α.

seq) in primary intestinal epithelial monolayers. Single-cell resolution enabled detailed kinetic profiling of dynamic transcriptional states that correspond to the early acute UPR followed by adaptive restoration of ER homeostasis or terminal cell death. In parallel, we performed a genome-wide CRISPR screen to identify regulatory nodes that control the terminal UPR. Integrative analysis of CRISPR screen results

with single-cell transcriptional profiling identified QRICH1 as a critical determinant of cellular entry into the terminal versus adaptive UPR. We demonstrate that QRICH1 is a key effector of the PERK-eIF2α axis of the UPR and that its translation is regulated by an upstream open reading frame in the QRICH1 mRNA. Using a combination of RNA-seq and chromatin immunoprecipitation sequencing (ChIP-seq), we show that QRICH1 bound promoter regions to control a transcriptional module that regulates protein translation and secretory networks. QRICH1-mediated translational activation increased protein flux into the ER and proteotoxicity, whereas QRICH1 knockout protected intestinal epithelial cells from proteotoxicity. Finally, to assess the role of QRICH1 in human disease, we analyzed biopsies of patients with ulcerative colitis (UC) and found evidence of enrichment of the QRICH1 transcriptional signature in inflamed colon biopsies, particularly in secretory epithelial cells and enterocytes. The QRICH1 transcriptional signature was also up-regulated in biopsies of patients with nonalcoholic steatohepatitis and in inflamed and cirrhotic samples from liver biopsies.

**CONCLUSION:** Here, we identify a distinct arm of the PERK-eIF2α axis mediated by the transcriptional regulator QRICH1. Cells dynamically respond to ER stress by inducing up-regulation of QRICH1, which modulates translation and transit of proteins through the ER-Golgi secretory pathway. Thus, QRICH1 acts as a regulator of a distinct transcriptional module that coordinates cellular stress responses to regulate protein synthesis and secretion under homeostatic and pathological conditions. Taken together, these findings suggest a broadly conserved role for the QRICH1 transcriptional program in managing cell stress responses and acting as a gatekeeper for controlling cellular entry into the adaptive versus terminal UPR. Mechanistic characterization of QRICH1 within this context provides insight into how cells manage responses to stress and expands our understanding of the UPR pathway, broadening our understanding of the molecular mechanisms by which cellular stress responses are dynamically regulated. ■

The list of author affiliations is available in the full article online.  
\*Corresponding author. Email: dgraham@broadinstitute.org (D.B.G.); aregev@broadinstitute.org (A.R.); xavier@molbio.mgh.harvard.edu (R.J.X.)  
Cite this article as K. You et al., *Science* 371, eabb6896 (2021). DOI: 10.1126/science.abb6896

**READ THE FULL ARTICLE AT**  
<https://doi.org/10.1126/science.abb6896>



## RESEARCH ARTICLES

## BATTERIES

## A rechargeable zinc-air battery based on zinc peroxide chemistry

Wei Sun<sup>1</sup>, Fei Wang<sup>2\*</sup>, Bao Zhang<sup>3</sup>, Mengyi Zhang<sup>1</sup>, Verena Küpers<sup>1</sup>, Xiao Ji<sup>4</sup>, Claudia Theile<sup>1</sup>, Peter Bieker<sup>1</sup>, Kang Xu<sup>5\*</sup>, Chunsheng Wang<sup>4,6\*</sup>, Martin Winter<sup>1,7\*</sup>

Rechargeable alkaline zinc-air batteries promise high energy density and safety but suffer from the sluggish 4 electron ( $e^-$ )/oxygen ( $O_2$ ) chemistry that requires participation of water and from the electrochemical irreversibility originating from parasitic reactions caused by caustic electrolytes and atmospheric carbon dioxide. Here, we report a zinc- $O_2$ /zinc peroxide ( $ZnO_2$ ) chemistry that proceeds through a  $2e^-/O_2$  process in nonalkaline aqueous electrolytes, which enables highly reversible redox reactions in zinc-air batteries. This  $ZnO_2$  chemistry was made possible by a water-poor and zinc ion ( $Zn^{2+}$ )-rich inner Helmholtz layer on the air cathode caused by the hydrophobic trifluoromethanesulfonate anions. The nonalkaline zinc-air battery thus constructed not only tolerates stable operations in ambient air but also exhibits substantially better reversibility than its alkaline counterpart.

**M**etal-air batteries provide tantalizing solutions to the next-generation energy storage systems (1–3), among which zinc-air batteries (ZABs) are of interest for their potential low cost, high safety, environmental friendliness, and high energy density (4). Most previous efforts focus on a Zn anode and a porous carbonaceous air cathode in alkaline aqueous electrolytes (5). As a nonrechargeable chemistry, primary alkaline ZABs have been known since the 19th century, with successful commercial products in medical and telecommunication applications, such as miniature hearing aids and wireless messaging devices (6, 7). However, the key challenge for a reversible ZAB comes from the strongly alkaline electrolyte, which is chemically unstable toward the active cathode material (ambient air) and, to a large extent, causes electrochemical irreversibility at the Zn metal anode (8). Zn metal in an alkaline environment suffers from formation of dendrites with high surface area, well known from Li metal anodes (9, 10), nonuniform electrodeposition and electrodisolution, and persistent corrosion that consumes electrolyte (11). To counter this, a considerable excess of Zn

has to be used, leading to substantial underutilization of its theoretical capacity (12). On the cathode side, the reaction between alkaline electrolytes and  $CO_2$  in air produces insoluble carbonate salts, which irreversibly consumes electrolyte and also physically clogs and chemically deactivates the porous air cathode (13). Hence, neat  $O_2$  atmosphere instead of ambient air is usually required (14, 15), inducing complicated cell designs and a further reduction in energy density. Because the redox reaction of  $O_2$  occurs through a sluggish 4 electron ( $e^-$ ) O–O bond cleavage and formation pathway in conventional alkaline electrolytes, bifunctional catalysts have to be used on the cathode (16).

Efforts to improve the reversibility of ZABs have focused on developing bifunctional catalysts for the air cathode (17) or prolonging the Zn anode life span by electrode architecture design or electrolyte additives (11, 18). Recent efforts have demonstrated that near-neutral aqueous electrolytes could suppress the formation of Zn dendrite and carbonates (19, 20). Highly reversible Zn electrodeposition and electrodisolution were also achieved in certain superconcentrated electrolytes (12). However, the core challenge presented by a sluggish  $4e^-$  pathway oxygen reduction reaction (ORR) chemistry at the cathode remained unexplored.

In most aqueous metal-air batteries using alkaline electrolytes and metals such as Zn, magnesium, iron, and aluminum,  $H_2O$  always participates in the ORR reaction through a  $4e^-$  pathway producing  $OH^-$  (21), which constitutes the root of the poor reversibility. By contrast, an alternative  $2e^-$  ORR reaction has been known as the dominating pathway in certain nonaqueous metal-air batteries, such as  $Li-O_2$  batteries with  $Li_2O_2$  as a product

(22), in which the  $4e^-$  ORR is suppressed. Superconcentrated electrolytes such as “water-in-salt” electrolytes provide a transition region between aqueous and nonaqueous realms. There have been reports showing that an aprotic ORR reaction could occur in water-in-salt electrolytes (23), in which the  $H_2O$  was excluded from the primary solvation structures of the cations (24). By comparing the hydrophobicity of three anions of Zn salts (fig. S1), we selected the hydrophobic trifluoromethanesulfonate ( $OTf^-$ ) anion with a large size as a constituent of the electrolyte solute. Because hydrophobic  $OTf^-$  anion will absorb on the cathode surface because of the electrostatic force, a localized  $H_2O$ -poor environment will be created in the inner Helmholtz layer (IHL) (25), enabling the aprotic  $2e^-$  ORR reaction in a dilute aqueous electrolyte.

## Zn-air cells in nonalkaline electrolytes

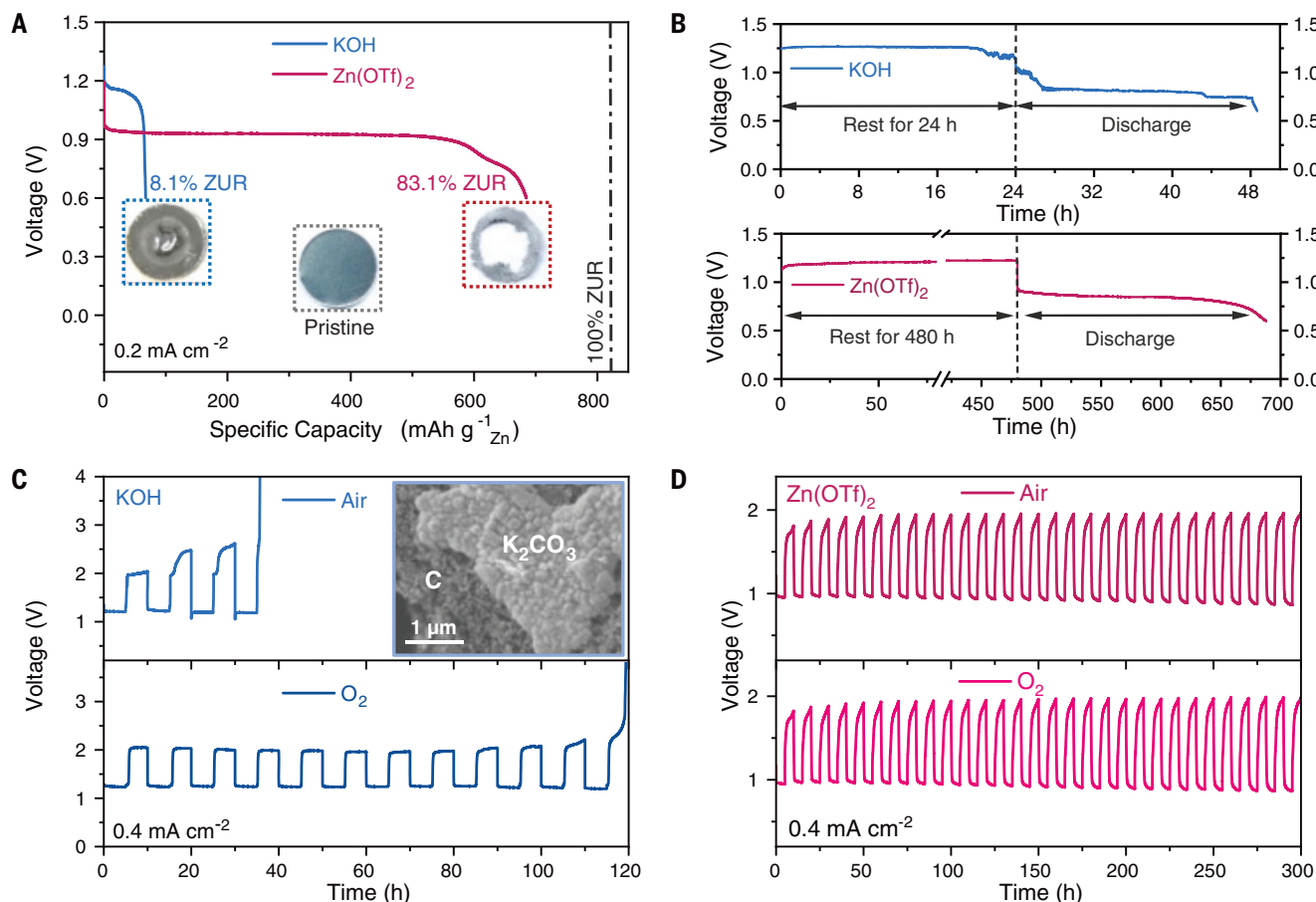
Ambient air was used as the cathode active agent in all Zn-air cell configurations (fig. S2). Nonalkaline electrolyte of  $1\text{ mol kg}^{-1} Zn(OTf)_2$  presents a well-defined discharge plateau at  $\sim 1.0\text{ V}$  with an areal capacity of  $52\text{ mA}\cdot\text{hour cm}^{-2}$ , corresponding to a specific capacity of  $684\text{ mA}\cdot\text{hour g}^{-1}$  (based on Zn anode) and a Zn utilization ratio (ZUR) of 83.1% (Fig. 1A). By contrast, only an 8.1% ZUR was achieved in alkaline electrolyte consisting of  $6\text{ mol kg}^{-1} KOH$ . The photographs of pristine and cycled Zn anodes in different electrolytes confirm the high ZUR in the  $Zn(OTf)_2$  electrolyte, in which most of the Zn foil dissolved. At higher rates of 0.3 and  $1\text{ mA cm}^{-2}$ , 70.8 and 52.9% ZUR could be achieved (fig. S3A). Replacing the Zn foil with Zn powder further increased the ZUR to 93.7% (fig. S3B).

After resting in the KOH electrolyte for 24 hours, the open circuit voltage (OCV) of the Zn-air cell starts fluctuating, followed by a lower discharge plateau and capacity fading (Fig. 1B), indicating the poor storage and discharge performance of the Zn-air cell with KOH electrolyte. By contrast, a rather stable OCV persists in the  $Zn(OTf)_2$  electrolyte for 480 hours, followed by a flat discharge plateau at  $\sim 1.0\text{ V}$ , indicating stable discharge. The  $Zn(OTf)_2$  electrolyte also enabled a more stable performance in an on-off discharge procedure that emulated a practical duty cycle (fig. S4, A and B). Persistent corrosion of the Zn anode in KOH electrolyte (26) can be confirmed by x-ray diffraction (XRD) patterns and photographs (fig. S4C), whereas Zn anodes in  $Zn(OTf)_2$  electrolyte basically maintain the pristine surface. Zn reversibility was further investigated using Zn||Zn symmetrical cells (fig. S5), in which the voltage profiles of repeated electrodeposition and electrodisolution of Zn metal anode in the  $Zn(OTf)_2$  electrolyte remained stable after more than 150 hours (75 cycles),

<sup>1</sup>MEET Battery Research Center, Institute of Physical Chemistry, University of Münster, Münster, Germany.

<sup>2</sup>Department of Materials Science, Fudan University, Shanghai, China. <sup>3</sup>School of Optical and Electronic Information, Huazhong University of Science and Technology, Wuhan, Hubei, China. <sup>4</sup>Department of Chemical and Biomolecular Engineering, University of Maryland, College Park, MD, USA. <sup>5</sup>Energy Storage Branch, Biomaterials and Energy Division, Sensor and Electro Devices Directorate, U.S. Army Research Laboratory, Adelphi, MD, USA. <sup>6</sup>Department of Chemistry and Biochemistry, University of Maryland, College Park, MD, USA. <sup>7</sup>Helmholtz Institute Münster, IEK-12, Forschungszentrum Jülich GmbH, Münster, Germany.

\*Corresponding author. Email: martin.winter@uni-muenster.de (M.W.); cswang@umd.edu (C.W.); conrad.k.xu.civ@mail.mil (K.X.); feiw@fudan.edu.cn (F.W.)



**Fig. 1. Electrochemical performance of Zn-air cells in a  $\text{Zn}(\text{OTf})_2$  electrolyte ( $1 \text{ mol kg}^{-1}$ ) and in a conventional KOH electrolyte ( $6 \text{ mol kg}^{-1}$ ).** (A) Galvanostatic discharge profiles of Zn-air cells in KOH (blue) and  $\text{Zn}(\text{OTf})_2$  (red) electrolytes at  $0.2 \text{ mA cm}^{-2}$  (cutoff voltage:  $0.6 \text{ V}$ ). The corresponding ZURs were indexed for comparison. Insets are photographs of the pristine Zn anode (middle), the Zn anode after discharge in KOH (left), and  $\text{Zn}(\text{OTf})_2$  (right) electrolytes. (B) Storage performance of Zn-air cells using KOH (24-hour rest before discharge) and  $\text{Zn}(\text{OTf})_2$  (480-hour rest before discharge) electrolytes. (C and D) Galvanostatic discharge and charge profiles of Zn-air cells in (C) KOH and (D)  $\text{Zn}(\text{OTf})_2$  electrolytes in a capacity fixed mode (fixed capacity:  $2 \text{ mA-hour cm}^{-2}$ ) at  $0.4 \text{ mA cm}^{-2}$  under air and  $\text{O}_2$  atmospheres, respectively. Inset is an SEM image of the air cathode after two cycles in KOH electrolyte under air atmosphere.

whereas a sudden increase in polarization occurred within only six cycles in the KOH electrolyte, indicating Zn dendrite formation. Using XRD and scanning electron microscopy (SEM), after only two cycles, a ZnO passivation layer was detected on the Zn anode recovered from the KOH electrolyte, whereas no new reflections appeared on Zn recovered from  $\text{Zn}(\text{OTf})_2$  electrolyte (fig. S6A). After cycling, a considerable morphology change was observed on Zn recovered from the KOH electrolyte (fig. S6, B and C), whereas the Zn deposition morphology remained relatively uniform in the  $\text{Zn}(\text{OTf})_2$  electrolyte (fig. S6D).

On the cathode side,  $\text{K}_2\text{CO}_3$  was observed after only two cycles in the KOH electrolyte, apparently generated by the reaction between KOH and the  $\text{CO}_2$  in air (fig. S7A).  $\text{K}_2\text{CO}_3$  precipitation clogged the porous structure of the air cathode and insulated the electron pathways (inset of Fig. 1C and fig. S7, B and C), contributing to severe loss of cathode activity

and capacity fading. The separator also turned brown after cycling in the KOH electrolyte, indicating carbon corrosion (27) (fig. S7E). In comparison, no apparent change could be found at both the air cathode and separator after cycling in the  $\text{Zn}(\text{OTf})_2$  electrolyte (fig. S7). As a result, the Zn-air cell in the KOH electrolyte lasted for only three cycles. Replacing air with neat  $\text{O}_2$  extended the life of such a cell to more than 12 cycles (Fig. 1C). When resting in neat  $\text{CO}_2$ , the OCV of the Zn-air cell started fluctuating immediately (fig. S8), showing the sensitivity of alkaline electrolyte toward  $\text{CO}_2$ . In the  $\text{Zn}(\text{OTf})_2$  electrolyte, however, the cells delivered similar performance in air and neat  $\text{O}_2$  (Fig. 1D). To further confirm the reproducibility, two identical cells were constructed and then cycled under either air or neat  $\text{O}_2$ , and the identical behavior verified the air tolerance of  $\text{Zn}(\text{OTf})_2$  electrolytes (fig. S9).

Besides  $\text{Zn}(\text{OTf})_2$  electrolytes, a  $\text{ZnSO}_4$ -based electrolyte was considered as another non-

alkaline alternative to the KOH electrolyte with better electrochemical reversibility (28–30). Using a pH-monitoring cell (fig. S2B), we tracked the electrolyte pH value for both  $\text{Zn}(\text{OTf})_2$  and  $\text{ZnSO}_4$  during operation of the Zn-air cells. The  $\text{Zn}(\text{OTf})_2$  electrolyte maintained a stable pH in both the discharge (ORR) and the charge [oxygen evolution reaction (OER)] processes, but the pH of the  $\text{ZnSO}_4$  electrolyte increased steadily during discharge (ORR) and decreased during charge (OER), apparently as a result of  $\text{OH}^-$  generation during ORR and consumption during OER (Fig. 2A). In extended cycling, the  $\text{Zn}(\text{OTf})_2$  electrolyte maintained a stable pH value (fig. S10).

To identify the discharge products, air cathodes were disassembled from the discharged Zn-air cells. Additional XRD patterns were indexed to  $\text{ZnO}_2$  on the cathode recovered from the  $\text{Zn}(\text{OTf})_2$  electrolytes (Fig. 2B). The absence of any other diffraction reflections except  $\text{ZnO}_2$  indicated that no other well-crystallized

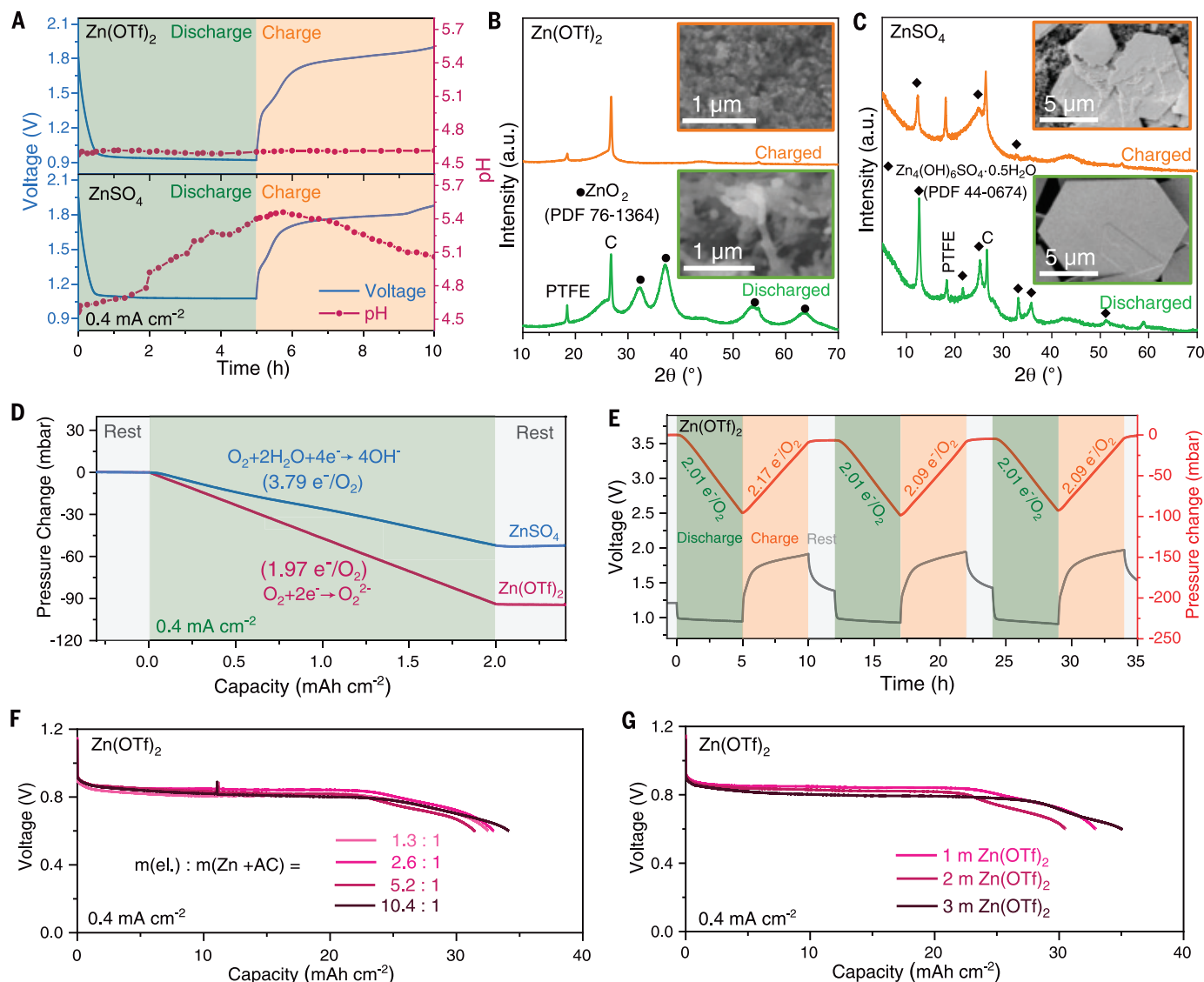


products were formed, which is further confirmed by energy-dispersive x-ray mapping, Raman, and OIs x-ray photoelectron spectroscopy, respectively (figs. S11 and S12). As revealed by SEM (inset of Fig. 2B and fig. S11B), the formed  $\text{ZnO}_2$  had a fiber structure within the scale of a few hundred nanometers. Upon recharge,  $\text{ZnO}_2$  completely disappeared from XRD, SEM, and Raman spectra (Fig. 2B and figs. S11C and S12B), demonstrating its electrochemical reversibility. Combined with ex situ XRD at different states of discharge and charge (fig. S12, C and D), fully reversible  $\text{ZnO}_2$  formation and decomposition chemistry can be

established during cycling in the  $\text{Zn}(\text{OTf})_2$  electrolyte. The chemical stability of  $\text{ZnO}_2$  in  $\text{Zn}(\text{OTf})_2$  electrolyte as well as in neat water was also evidenced by XRD that showed no major changes during 30 days of storage (fig. S13).

In  $\text{ZnSO}_4$  electrolyte, zinc sulfate hydroxide [ $\text{Zn}_4(\text{OH})_6\text{SO}_4 \cdot 0.5\text{H}_2\text{O}$  (ZHS)] was identified as a discharge product (Fig. 2C and fig. S14, A to D). The formed ZHS had a typical large flake-like structure at the scale of tens of micrometers, consistent with (31, 32). Upon charging, part of the insulating ZHS flakes remained, because of its poor reversibility in the  $\text{ZnSO}_4$  electrolyte (fig. S14E).

The difference in electrochemical behavior caused by these two electrolytes could be attributed to the respective reaction mechanisms. The number of  $\text{e}^-$  ( $Z$ ) transferred to  $\text{O}_2$  during ORR or OER was determined by measuring the  $\text{O}_2$  consumption or evolution and the corresponding charge transfer based on the ideal gas law and Faraday's law (33). The pressure decrease of a  $\text{Zn-O}_2$  cell in  $\text{Zn}(\text{OTf})_2$  was about two times that in the  $\text{ZnSO}_4$  electrolyte, corresponding to a  $1.97\text{e}^-$  transfer per  $\text{O}_2$  molecule ( $\text{e}^-/\text{O}_2$ ) in  $\text{Zn}(\text{OTf})_2$  and a  $3.79\text{e}^-/\text{O}_2$  in  $\text{ZnSO}_4$  (Fig. 2D and fig. S15). The pressure of the  $\text{Zn-O}_2$  cell using  $\text{Zn}(\text{OTf})_2$  electrolyte is

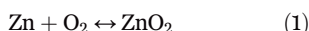


**Fig. 2. Elucidating electrochemical reaction mechanisms of Zn-air cells in  $\text{Zn}(\text{OTf})_2$  ( $1\text{ mol kg}^{-1}$ ) and  $\text{ZnSO}_4$  ( $1\text{ mol kg}^{-1}$ ) electrolytes.** (A) Typical second-cycle galvanostatic discharge and charge profiles of Zn-air cells using  $\text{Zn}(\text{OTf})_2$  and  $\text{ZnSO}_4$  electrolytes at  $0.4\text{ mA cm}^{-2}$  and the corresponding recorded electrolyte pH values. (B and C) XRD patterns and SEM images of air cathodes obtained after discharge and recharge in (B)  $\text{Zn}(\text{OTf})_2$  and (C)  $\text{ZnSO}_4$  electrolytes, respectively. a.u., arbitrary units; C, carbon; PDF, powder diffraction file; PTFE, polytetrafluoroethylene. (D) Pressure change in

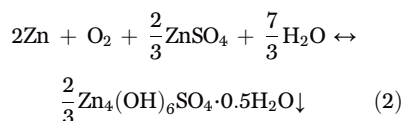
the gas reservoir of  $\text{Zn-O}_2$  cells using  $\text{Zn}(\text{OTf})_2$  and  $\text{ZnSO}_4$  electrolytes during a discharge process under neat  $\text{O}_2$  atmosphere. (E) Pressure change of  $\text{Zn-O}_2$  cells using  $\text{Zn}(\text{OTf})_2$  electrolyte during galvanostatic discharge and charge at  $0.4\text{ mA cm}^{-2}$ . (F and G) Typical galvanostatic discharge profiles of Zn-air cells using  $\text{Zn}(\text{OTf})_2$  electrolyte (F) with varying mass ratios of electrolyte to electrodes [ $\text{Zn}$  anode and air cathode,  $m(\text{el.}):m(\text{Zn} + \text{AC})$ ] and (G) with varying salt concentrations at  $0.4\text{ mA cm}^{-2}$ . AC, air cathode; el., electrolyte; m,  $\text{mol kg}^{-1}$ .

reversible, as shown in Fig. 2E, where the pressure increased back to the starting value when a discharge and charge cycle finished, confirming the reversible  $O_2$  consumption and evolution during the cycling. The Z values in multiple discharging processes are all close to  $2e^-/O_2$ , corresponding to the formation of  $ZnO_2$ . However, Z values in these charging processes are slightly higher than 2, which might originate from minor  $H_2O$  decomposition with a  $4e^-/O_2$  pathway.

Combining the aforementioned findings about the discharge product and pH value during the discharge and charge process, the overall reaction of the ZAB chemistry in  $Zn(OTf)_2$  can be expressed as reversible  $ZnO_2$  formation and decomposition (Eq. 1)



whereas the cell reaction in  $ZnSO_4$  electrolytes is reversible ZHS formation and decomposition (Eq. 2)

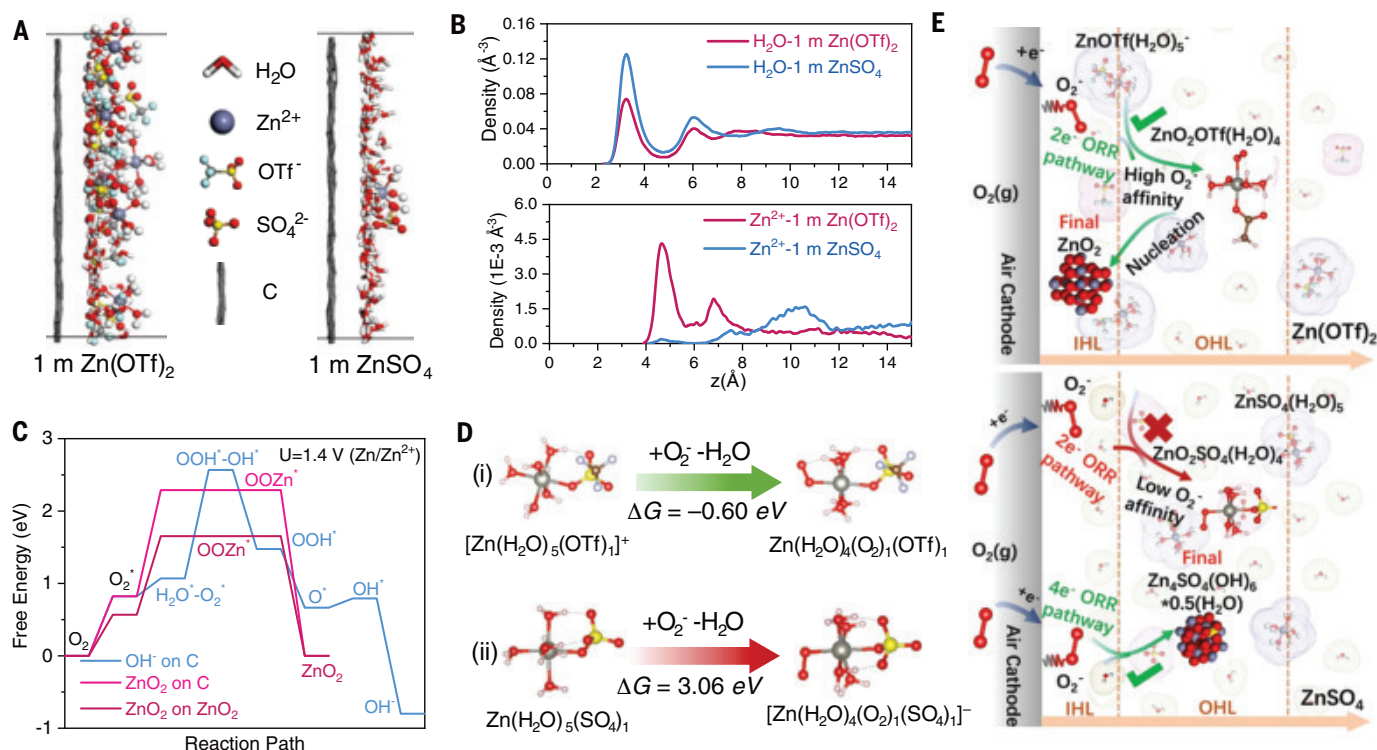


The ORR reaction in the  $ZnSO_4$  electrolyte generates  $OH^-$ , similar to that in conventional alkaline (KOH) electrolytes (34, 35).  $ZnSO_4$  will subsequently react with  $OH^-$ , creating a complex hydrate ZHS (37). In other words, ZHS in this case is limited by the availability of  $ZnSO_4$  or  $H_2O$  from the electrolyte, in contrast with the cathode ORR reaction in the  $Zn(OTf)_2$  electrolyte, in which no electrolyte should be consumed. The reversible  $ZnO_2$  formation and decomposition mechanism resembles that involving  $Li_2O_2$  formation at the cathode in nonaqueous Li- $O_2$  batteries (36, 37). The two different mechanisms thus lead to substantially different dependences of the cell capacity on the amount of electrolyte. The capacity delivered in  $ZnSO_4$  electrolytes is restricted by the electrolyte amount (fig. S16), whose steady consumption will also induce a drop in ion conductivity; hence, excessive electrolyte is needed to maintain a discharge and charge reaction over cycling, at the expense of practical energy density. When using  $Zn(OTf)_2$  electrolytes with different electrolyte/electrode mass ratios, essentially identical discharge performance was obtained (Fig. 2F). Similar discharge curves were also demonstrated in the concentra-

tion range of  $Zn(OTf)_2$  from 1 to 3 mol  $kg^{-1}$  (Fig. 2G and fig. S17). Therefore, the  $2e^-/O_2$  reaction mechanism enabled by  $Zn(OTf)_2$  electrolyte not only offers a reversible ZAB chemistry but also makes it possible to construct high-energy density ZABs with lean electrolyte volume. An alternative zinc salt with larger fluorinated anion, zinc bis(trifluoromethanesulfonyl)imide [ $Zn(TFSI)_2$ ], was also evaluated. Similar electrochemical performance and cell reaction products were observed with 1 mol  $kg^{-1}$   $Zn(TFSI)_2$  (fig. S18), verifying that “superconcentration” is not needed as long as a proper anion is selected.

### $ZnO_2$ formation mechanism

Multiscale simulations were used to understand the origin for the different mechanisms observed in different Zn-ion electrolytes. Molecular dynamics simulations showed that the bulk liquid structures in 1 mol  $kg^{-1}$   $Zn(OTf)_2$  and  $ZnSO_4$  electrolytes were similar, in which  $Zn^{2+}$  was coordinated with about six  $H_2O$  molecules (fig. S19, A to D). However, a major difference existed between the hydrophobicity of  $SO_4^{2-}$  and  $OTf^-$  owing to the strongly hydrophobic  $CF_3$  group on  $OTf^-$  (fig. S19, E and F). This difference is reflected in the fact that



**Fig. 3. ORR mechanism in  $Zn(OTf)_2$  and  $ZnSO_4$  electrolytes.** (A) Schematic snapshots (side view) of the interfacial structures at the air cathodes in  $Zn(OTf)_2$  and  $ZnSO_4$  electrolytes with positive applied potential ( $U = 1.4$  V versus  $Zn/Zn^{2+}$ ). (B) Corresponding interfacial cumulated density profiles of the  $H_2O$  and  $Zn^{2+}$  in the  $Zn(OTf)_2$  and  $ZnSO_4$  electrolytes, respectively. z, distance from the cathode surface. (C) Free-energy diagrams of ORR processes with a

reaction path through  $OH^-$  on the C surface, through  $ZnO_2$  formation on the C surface, and through  $ZnO_2$  formation on the  $ZnO_2$  surface ( $U = 1.4$  V versus  $Zn/Zn^{2+}$ ). (D) Desolvation and superoxide-containing ion pair formation process in (i)  $Zn(OTf)_2$  and (ii)  $ZnSO_4$  electrolytes. (E) Schematic illustration of reaction processes in the IHL and outer Helmholtz layer (OHL) at the surface of the air cathode in  $Zn(OTf)_2$  and  $ZnSO_4$  electrolytes, respectively.



commercial  $\text{Zn}(\text{OTf})_2$  salt is normally available as anhydrous powder, whereas  $\text{ZnSO}_4$  salt usually comes in the form of hydrated crystals (fig. S1). Such hydrophobicity disparity directly determined the behavior of these anions when assembling at the air cathode surface, leading to distinctly different electrochemical double-layer structures. At an applied potential of  $U = 0.8 \text{ V}$  versus  $\text{Zn}/\text{Zn}^{2+}$  (simulated potential of zero charge), the hydrophobic  $\text{OTf}^-$  anions created an  $\text{H}_2\text{O}$ -poor environment, whereas the hydrophilic  $\text{SO}_4^{2-}$  anion led to an  $\text{H}_2\text{O}$ -rich environment at the IHL (fig. S20, A to C). Positive deviation of the applied potential toward  $U = 1.4 \text{ V}$  versus  $\text{Zn}/\text{Zn}^{2+}$  (simulated potential of positive charge) further enriched the population of the hydrophobic  $\text{OTf}^-$  anions in the IHL of the cathode, which simultaneously decreased the  $\text{H}_2\text{O}$  presence and reduced the opportunity of  $\text{H}_2\text{O}$ -related reactions (Fig. 3A and fig. S20, D and E). By contrast, the IHL- $\text{H}_2\text{O}$  molecular ratio experienced little change in the  $\text{ZnSO}_4$  electrolyte as the potential shifted. Most of the  $\text{Zn}^{2+}$  ions were expelled from the IHL in the  $\text{ZnSO}_4$  electrolyte, whereas abundant  $\text{Zn}^{2+}$  ions in the  $\text{Zn}(\text{OTf})_2$  electrolyte remained within a distance of  $6 \text{ \AA}$  from the cathode. Such interfacially accumulated density profiles of  $\text{Zn}^{2+}$  and  $\text{H}_2\text{O}$  indicated that more  $\text{Zn}^{2+}$  but less  $\text{H}_2\text{O}$  molecules assembled near the air cathode in the  $\text{Zn}(\text{OTf})_2$  electro-

lyte than in the  $\text{ZnSO}_4$  electrolyte (Fig. 3B). This relatively  $\text{H}_2\text{O}$ -poor and  $\text{Zn}^{2+}$ -rich IHL environment created by a  $\text{Zn}(\text{OTf})_2$  electrolyte reduced the opportunity of  $\text{H}_2\text{O}$ -related reactions but provided better access to  $\text{Zn}^{2+}$  for the ORR process.

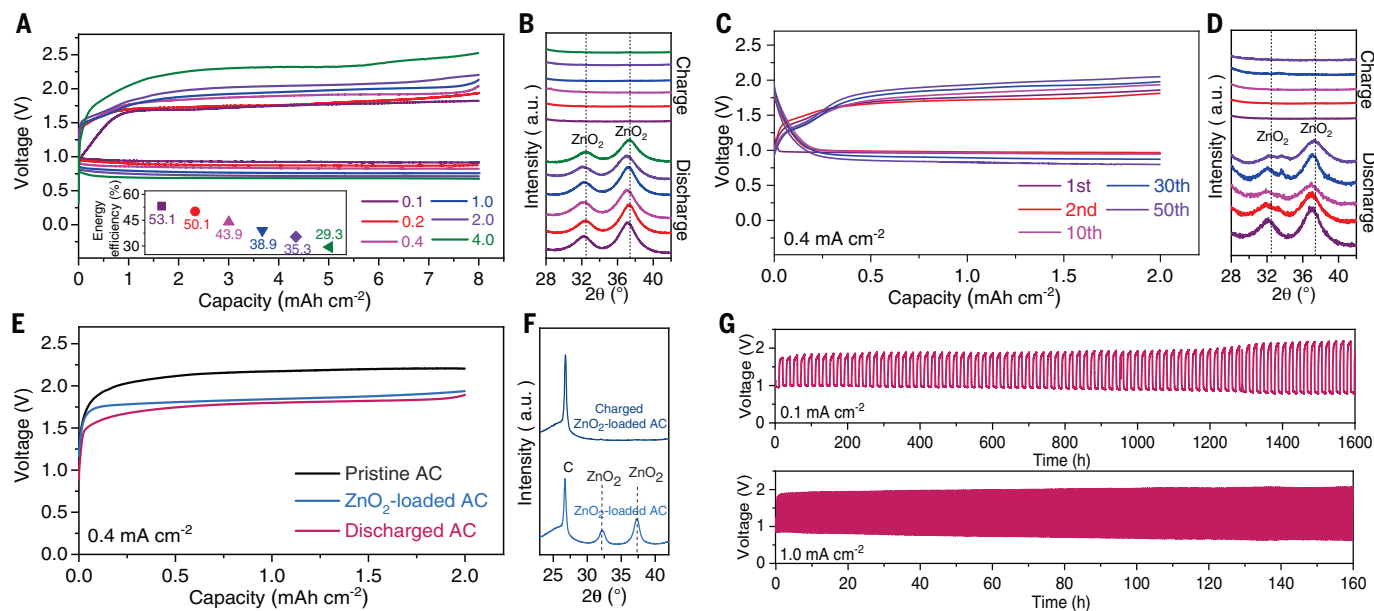
Density functional theory calculations were also performed to predict possible ORR pathways. Two reaction mechanisms ( $2e^-$  and  $4e^-$  transfer) were compared for reactions involving both  $\text{Zn}^{2+}$  and  $\text{H}_2\text{O}$  (fig. S21). After excluding those with higher free-energy barriers, the two most probable pathways, i.e., the  $2e^-$  transfer with  $\text{Zn}^{2+}$  and the  $4e^-$  transfer with  $\text{H}_2\text{O}$  on the carbon surface, remained (Fig. 3C). The overpotential of the  $\text{Zn}^{2+}$ -involving  $2e^-$  mechanism was slightly lower than that of the  $\text{H}_2\text{O}$ -involving  $4e^-$  mechanism, mainly because of the sluggish and difficult H-OH bond cleaving process. Hence, the  $4e^-$   $\text{H}_2\text{O}$  decomposition pathway should be disfavored in  $\text{Zn}(\text{OTf})_2$  electrolytes. As  $\text{ZnO}_2$  formation on its own surface would encounter a much lower free-energy barrier,  $\text{ZnO}_2$  growth should follow an accelerated process once overcoming the initial nucleation stage, which is in good agreement with the SEM images after different discharge times (fig. S22).

Quantum chemistry calculations reveal that the weak affinity between  $\text{Zn}^{2+}$  cations and  $\text{OTf}^-$  anions in  $\text{Zn}(\text{OTf})_2$  electrolyte favors the formation of superoxide-containing ion

pairs and advances the subsequent  $2e^-$  pathway mechanism involving  $\text{Zn}^{2+}$ , whereas the higher affinity between  $\text{Zn}^{2+}$  cations and  $\text{SO}_4^{2-}$  anions inhibits  $\text{Zn}^{2+}$  desolvation and the subsequent formation process of superoxide-containing ion pairs (Fig. 3D and fig. S23). The predicted ORR processes inside the electrochemical double layer near the cathode surface in different electrolytes are summarized in Fig. 3E, where an  $\text{H}_2\text{O}$ -poor and  $\text{Zn}^{2+}$ -rich structure brought in by the hydrophobic nature of the  $\text{OTf}^-$  anions and the weak affinity between  $\text{Zn}^{2+}$  cations and  $\text{OTf}^-$  anions are the two main contributors to the preferred  $2e^-$  ORR chemistry, leading to the highly reversible  $\text{ZnO}_2$  formation, which is impossible in electrolytes using highly hydrophilic anions such as  $\text{SO}_4^{2-}$ .

### Highly reversible Zn-air cells in $\text{Zn}(\text{OTf})_2$ electrolytes

The rate performance of the Zn-air cells was examined at different current densities with a fixed capacity of  $8 \text{ mA-hour cm}^{-2}$  (Fig. 4A). Corresponding XRD patterns of air cathodes after discharging and charging confirmed reversible  $\text{ZnO}_2$  formation and oxidation at different current densities (Fig. 4B and fig. S24). A slightly increasing polarization was observed when the current increased from  $0.1$  to  $2.0 \text{ mA cm}^{-2}$ . At a high current of  $4.0 \text{ mA cm}^{-2}$ , the large overpotential could



**Fig. 4. Cycling stability of Zn-air cells based on  $\text{ZnO}_2$ .** (A) Galvanostatic voltage profiles and energy efficiencies (inset) of the Zn-air cells in a capacity fixed mode (fixed capacity:  $8.0 \text{ mA-hour cm}^{-2}$ ) at current densities of  $0.1$ ,  $0.2$ ,  $0.4$ ,  $1.0$ ,  $2.0$ , and  $4.0 \text{ mA cm}^{-2}$ . (B) Corresponding XRD patterns of air cathodes obtained after discharge and charge at different current densities. (C) Galvanostatic voltage profiles of the Zn-air cell in a capacity fixed mode (fixed

capacity:  $2.0 \text{ mA-hour cm}^{-2}$ ) at a current density of  $0.4 \text{ mA cm}^{-2}$ . (D) Corresponding XRD patterns of air cathodes obtained after different cycles. (E) Galvanostatic charge profiles of the pristine air cathode, the  $\text{ZnO}_2$ -loaded air cathode, and the discharged air cathode. (F) Corresponding XRD patterns of the  $\text{ZnO}_2$ -loaded air cathodes before and after charge. (G) Long-term cycling performance of Zn-air cells at current densities of  $0.1$  and  $1.0 \text{ mA cm}^{-2}$ .

lead to oxidation of  $\text{H}_2\text{O}$ . Therefore, powerful catalysts for the  $2e^-$  ORR or OER reaction are still needed for better energy efficiency. Without catalysts, the round-trip energy efficiencies (<55%, inset of Fig. 4A) of the Zn-air cell are relatively low when compared with those of Li-ion battery technology (>90%) (38, 39). However, the overvoltage observed for this nonalkaline ZAB is already comparable to many alkaline ZABs and Li- $\text{O}_2$  batteries using catalysts (table S1). As demonstrated in the three-electrode cells, the high overvoltage of the Zn-air full cells mainly comes from the ORR and OER at the air cathode (fig. S25). One would infer that a proper bifunctional catalyst or redox mediator may further lower the overpotential at the cathode and improve the energy efficiency. Two approaches, carbon cathode modification and increased temperature, can be adopted to reduce the overpotential of air cathodes (fig. S26).

The Zn-air cell demonstrated no apparent polarization increase after 50 cycles at  $0.4 \text{ mA cm}^{-2}$  (Fig. 4C). XRD patterns during different cycles reveal that the  $\text{ZnO}_2$  formed during discharge fully vanished after the subsequent charge process (Fig. 4D and fig. S27), further corroborating the reversibility of the  $\text{O}_2/\text{ZnO}_2$  redox reaction. The electrochemical reversibility of  $\text{ZnO}_2$  was further confirmed by charging chemically synthesized  $\text{ZnO}_2$  in a similar air cell setup (Fig. 4E and fig. S28A). By loading the synthesized  $\text{ZnO}_2$  into an air cathode ( $\text{ZnO}_2$ -loaded AC), it delivered a similar charge behavior as the electrochemically formed  $\text{ZnO}_2$  in the air cathode (discharged AC), differing from the blank carbon cathode (pristine AC), in which only electrolyte decomposition occurred. XRD and SEM images revealed the disappearance of  $\text{ZnO}_2$  after charging (Fig. 4F and fig. S28, B to D), indicating the high electrochemical activity of  $\text{ZnO}_2$ . The gas compositions of the Zn-air cell after charging and overcharging were analyzed using a gas chromatography-thermal conductivity detector. Most of the gas was  $\text{O}_2$ , with no  $\text{CO}_2$  detected, again confirming the reversibility of  $\text{ZnO}_2$  but also indicating the absence of carbon corrosion in the air cathode (fig. S29).

The Zn-air cell operated stably for 1600 hours in ambient air at a current density of  $0.1 \text{ mA cm}^{-2}$  with a 10-hour charge and discharge duration per cycle (Fig. 4G). When cycled at a 10-times-higher current density of  $1.0 \text{ mA cm}^{-2}$ , a stable cycling performance for 160 hours (320 cycles) was achieved. The performance from this nascent chemistry is encouraging, but a practical device still needs more engineering optimization. As demonstrated in a three-electrode Zn-air cell configuration (fig. S30), the sluggish OER and ORR at the air cathode limits the kinetics of the Zn-air cell in the short term, and the instability of the Zn metal anode

eventually leads to cell failure. Moreover, because the ZAB in this work is designed to work in an open atmosphere, the electrolyte evaporation (water loss) becomes an inevitable issue, and a water management system might be needed to ensure a long-term practical operation (fig. S31). In nonalkaline electrolytes, the parasitic reactions of both the Zn anode and electrolytes could also present a challenge, owing to the  $\text{H}^+$  activity being higher than that in the alkaline counterpart (e.g.,  $6 \text{ mol kg}^{-1}$  KOH), and additional measures are needed to suppress the parasitic reactions as inspired by pioneering work (12, 40). Nevertheless, to demonstrate the simplicity of fabrication and applicability of this  $\text{ZnO}_2$  battery chemistry, a prototype square cell with more than 50 mA-hour energy storage was assembled and discharged at a large current of 20 mA (fig. S32), which displays a well-defined discharge plateau at  $\sim 1.0 \text{ V}$ , similar to that of the experimental Swagelok cell.

We demonstrated a nonalkaline ZAB based on a reversible  $\text{O}_2/\text{ZnO}_2$  chemistry. Comprehensive characterization and simulations identified the critical role of hydrophobic OTF<sup>-</sup> anions in dictating the electrochemical double-layer structure that favors the formation of  $\text{ZnO}_2$  and suppression of  $\text{H}_2\text{O}$ -involved reactions. Leveraging the high reversibility of both the air cathode and Zn metal anode in the  $\text{Zn}(\text{OTF})_2$  electrolyte, the Zn-air full cell demonstrated excellent cycling performance in ambient air despite a simple cell structure. Such tailoring of interfacial structures through electrolyte properties provides a solution to the electrochemical irreversibility that has been plaguing not only alkaline ZABs but also nearly all metal-air batteries for centuries, especially those with promising high theoretical energy densities using materials with abundance, but being only feasible in alkaline electrolytes as either primary or mechanically rechargeable batteries. Examples may include magnesium-air [theoretical specific energy:  $6815 \text{ watt-hours (Wh) kg}^{-1}$ ], iron-air ( $1229 \text{ Wh kg}^{-1}$ ), or aluminum-air ( $8076 \text{ Wh kg}^{-1}$ ) (38).

## REFERENCES AND NOTES

- C. Xia, C. Y. Kwok, L. F. Nazar, *Science* **361**, 777–781 (2018).
- G. Cong, W. Wang, N.-C. Lai, Z. Liang, Y.-C. Lu, *Nat. Mater.* **18**, 390–396 (2019).
- B. J. Hopkins, Y. Shao-Horn, D. P. Hart, *Science* **362**, 658–661 (2018).
- J. W. Choi, D. Aurbach, *Nat. Rev. Mater.* **1**, 16013 (2016).
- H.-F. Wang, Q. Xu, *Matter* **1**, 565–595 (2019).
- Y. Li, H. Dai, *Chem. Soc. Rev.* **43**, 5257–5275 (2014).
- D. Stock, S. Dongmo, J. Janek, D. Schröder, *ACS Energy Lett.* **4**, 1287–1300 (2019).
- A. R. Mainar et al., *Int. J. Energy Res.* **40**, 1032–1049 (2016).
- M. Winter, B. Barnett, K. Xu, *Chem. Rev.* **118**, 11433–11456 (2018).

- G. Bieker, M. Winter, P. Bieker, *Phys. Chem. Chem. Phys.* **17**, 8670–8679 (2015).
- D. E. Turney et al., *Chem. Mater.* **29**, 4819–4832 (2017).
- F. Wang et al., *Nat. Mater.* **17**, 543–549 (2018).
- J. Fu et al., *Adv. Mater.* **29**, 1604685 (2017).
- Y. Li et al., *Nat. Commun.* **4**, 1805 (2013).
- M. Prabu, K. Ketpang, S. Shanmugam, *Nanoscale* **6**, 3173–3181 (2014).
- J. Fu et al., *Adv. Mater.* **31**, 1805230 (2019).
- F. Cheng, J. Chen, *Chem. Soc. Rev.* **41**, 2172–2192 (2012).
- J. F. Parker et al., *Science* **356**, 415–418 (2017).
- F. W. Thomas Goh et al., *J. Electrochem. Soc.* **161**, A2080–A2086 (2014).
- C. Y. Chen, K. Matsumoto, K. Kubota, R. Hagiwara, Q. Xu, *Adv. Energy Mater.* **9**, 1900196 (2019).
- Z.-L. Wang, D. Xu, J.-J. Xu, X.-B. Zhang, *Chem. Soc. Rev.* **43**, 7746–7786 (2014).
- Z. Peng, S. A. Freunberger, Y. Chen, P. G. Bruce, *Science* **337**, 563–566 (2012).
- Q. Dong et al., *Chem* **4**, 1345–1358 (2018).
- L. Suo et al., *Science* **350**, 938–943 (2015).
- O. Borodin, J. Self, K. A. Persson, C. Wang, K. Xu, *Joule* **4**, 69–100 (2020).
- X. G. Zhang, *Corrosion and Electrochemistry of Zinc* (Plenum, 2013).
- P. N. Ross, H. Sokol, *J. Electrochem. Soc.* **131**, 1742–1750 (1984).
- D. Kundu, B. D. Adams, V. Duffort, S. H. Vajargah, L. F. Nazar, *Nat. Energy* **1**, 16119 (2016).
- W. Sun et al., *J. Am. Chem. Soc.* **139**, 9775–9778 (2017).
- F. Wang et al., *Energy Environ. Sci.* **11**, 3168–3175 (2018).
- H. Pan et al., *Nat. Energy* **1**, 16039 (2016).
- J. Huang et al., *Nat. Commun.* **9**, 2906 (2018).
- P. Hartmann et al., *Nat. Mater.* **12**, 228–232 (2013).
- E. Yeager, *Electrochim. Acta* **29**, 1527–1537 (1984).
- J. Suttivich et al., *Nat. Chem.* **3**, 546–550 (2011).
- P. G. Bruce, S. A. Freunberger, L. J. Hardwick, J.-M. Tarascon, *Nat. Mater.* **11**, 19–29 (2011).
- A. C. Luntz, B. D. McCloskey, *Chem. Rev.* **114**, 11721–11750 (2014).
- Y. Li, J. Lu, *ACS Energy Lett.* **2**, 1370–1377 (2017).
- P. Meister et al., *Chem. Mater.* **28**, 7203–7217 (2016).
- J. Y. Luo, W. J. Cui, P. He, Y. Y. Xia, *Nat. Chem.* **2**, 760–765 (2010).

## ACKNOWLEDGMENTS

We thank F. Horsthemke, F. J. Dohmann, and J. J. Jiang for technical support. **Funding:** This work at MEET was supported by the BMBF (Federal Ministry of Education and Research) within the projects “Mvbic” (03XP0249), “Melubatt” (03XP0110A), “MEET Hi-EnD III” (03XP0258A), and “MgMeAns” (03XP0140). At the University of Maryland, it was supported by the U.S. Department of Energy (DOE) (through ARPA-E grant DEAR0000389), and at the Army Research Laboratory, it was supported as part of the Joint Center for Energy Storage Research, an Energy Innovation Hub funded by the DOE, Office of Science, Basic Energy Sciences (through IAA SN2020957). **Author contributions:** W.S. and F.W. conceived and designed this work. M.W., C.W., and K.X. served as technical leads for this work. W.S., F.W., M.W., C.W., and K.X. contributed to the implementation and writing of the manuscript. Data collection and analysis were conducted by W.S. and F.W. Multiscale simulation was conducted by B.Z. and X.J. SEM and Raman characterizations were conducted by W.S., M.Z., and C.T. Energy density calculation was conducted by V.K. and P.B. **Competing interests:** The authors have no competing interests. **Data and materials availability:** All data are available in the manuscript or in the supplementary materials.

## SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/371/6524/46/suppl/DC1  
Materials and Methods  
Figs. S1 to S32  
Table S1  
References (41–68)

2 April 2020; accepted 19 November 2020  
10.1126/science.abb9554



## STEM CELLS

# Airway stem cells sense hypoxia and differentiate into protective solitary neuroendocrine cells

Manjunatha Shivaraju<sup>1,2,3</sup>, Udbhav K. Chitta<sup>4</sup>, Robert M. H. Grange<sup>5</sup>, Isha H. Jain<sup>6,7,8,9</sup>, Diane Capen<sup>10</sup>, Lan Liao<sup>11</sup>, Jianming Xu<sup>11</sup>, Fumito Ichinose<sup>5</sup>, Warren M. Zapol<sup>5</sup>, Vamsi K. Mootha<sup>6,7,8</sup>, Jayaraj Rajagopal<sup>1,2,3\*</sup>

Neuroendocrine (NE) cells are epithelial cells that possess many of the characteristics of neurons, including the presence of secretory vesicles and the ability to sense environmental stimuli. The normal physiologic functions of solitary airway NE cells remain a mystery. We show that mouse and human airway basal stem cells sense hypoxia. Hypoxia triggers the direct differentiation of these stem cells into solitary NE cells. Ablation of these solitary NE cells during hypoxia results in increased epithelial injury, whereas the administration of the NE cell peptide CGRP rescues this excess damage. Thus, we identify stem cells that directly sense hypoxia and respond by differentiating into solitary NE cells that secrete a protective peptide that mitigates hypoxic injury.

**A**irway neuroendocrine (NE) cells were first identified as epithelial cells that store and secrete amines and peptides from membrane-bound vesicles, mirroring the process of neurotransmitter release from neurons. Indeed, both neurons and airway NE cells secrete serotonin and calcitonin gene-related peptide (CGRP) (1, 2). NE cell hyperplasia occurs in diverse lung diseases, including neuroendocrine hyperplasia of infancy (NEHI), sudden infant death syndrome (SIDS), asthma, congenital pneumonia, pulmonary hypertension, cystic fibrosis, congenital diaphragmatic hernia, and chronic obstructive pulmonary disease (COPD) (3, 4). However, both the cause and relevance of NE cell hyperplasia remain unknown.

NE cells can be found as solitary cells or clustered as neuroepithelial bodies (NEBs). In mice, solitary pulmonary NE cells are found in the trachea and primary bronchi, but in humans, they are scattered throughout the airway tree. By contrast, the neuroepithelial bodies (NEBs) of mice are clustered groups of NE cells found specifically at airway branch points, whereas their presumed human NEB counterparts are less stereotypically distributed throughout the airway epithelium (5, 6).

NEBs are reported to perform various physiologic functions, including oxygen sensing (7), mechanotransduction (8, 9), modulating pulmonary blood flow (9), chemosensation (9), and regulating inflammation (10). Additionally, murine NEBs are thought to serve as niches for adjacent progenitor cells (11). However, the physiologic functions of solitary NE cells are largely unknown. One study of human solitary NE cells in vitro suggests that they function as airway chemosensory cells in vivo (12). Human solitary NE cells secrete CGRP in culture, and it is hypothesized that this neuropeptide links epithelial stimulus detection to the modulation of stem cell behavior (12).

Murine NEBs contain a subpopulation of NE stem cells that rarely produce new NE cells under steady-state conditions. However, following injury, NE stem cells generate more NE cells, leading to larger NEBs. The NE stem cells of NEBs are also plastic and contribute to the lineage of non-NE epithelial cells in their immediate vicinity (13). Despite this plasticity, genetic ablation of NEBs fails to elicit a gross effect on airway epithelial regeneration (14). By contrast, solitary NE cells display rapid turnover and are regularly replaced by basal stem cells (15, 16). The divergent lineage and kinetics of solitary NE cells and NEB-localized NE cells suggests that their functions may be distinct.

After birth, as the airway epithelium is first exposed to ambient oxygen, NE cell numbers decline (17–19). Hypoxia is known to delay the normal postnatal disappearance of NE cells in rabbit airway epithelia, but the mechanism is not understood (17). Elimination of oxygen-sensing prolyl hydroxylases (PHDs) from NEB-resident NE cells results in NEB growth (20). Because solitary NE cells are derived from basal stem cells rather than preexisting NE cells, we postulated that basal stem cells might possess an oxygen-sensing

mechanism that triggers hypoxia-stimulated stem cells to undergo solitary NE cell differentiation. We further speculated that these hypoxia-induced solitary NE cells would have a protective physiologic function in the setting of hypoxia. The coincident occurrence of solitary NE cells and basal cells is characteristic of the majority of the human airway tree, so we used the murine trachea as a model system to study solitary NE cell biology because it is the only part of the murine respiratory tree that contains both solitary NE cells and basal stem cells (21).

## Hypoxia induces solitary NE cell differentiation

When mice were exposed to hypoxic conditions (F<sub>2</sub>O<sub>2</sub> 8%) (fig. S1), we observed a significant time-dependent increase in solitary NE cells (Fig. 1, A and B). To exclude the possibility of injury-induced nonspecific expression of NE cell markers by non-NE cells, we assessed the expression of the NE cell fate-determining transcription factor *Ascl1*. *Ascl1-CreERT2::Rosa26-tdTomato* (hereafter referred to as *Ascl1-tdTomato*) mice were exposed to 20 days of hypoxia (fig. S2A), and tamoxifen was administered to label *Ascl1*<sup>+</sup> NE cells. We observed a significant increase in *Ascl1-tdTomato*<sup>+</sup> cells (fig. S2, B and C), confirming NE cell differentiation. Finally, quantifying the fraction of cells possessing NE cell-specific vesicles by electron microscopy confirmed the potent induction of bona fide solitary NE cells (Fig. 1, C and D).

We next sought to determine the cellular origin of the hypoxia-induced solitary NE cells. Because the deletion of PHDs, which mimics hypoxia, results in NEB growth (20), we assessed whether hypoxia could similarly cause solitary NE cells to self-renew. Preexisting solitary NE cells in *Ascl1-tdTomato* mice were lineage labeled, and then animals were subjected to hypoxia. The incorporation of 5-bromodeoxyuridine (BrdU) over the entire period of hypoxia was measured as an index of replication (fig. S3A). Although solitary NE cell numbers increased significantly after hypoxia (fig. S3, B and C), we did not observe any lineage-labeled cells or chromogranin A–positive (CHGA<sup>+</sup>) solitary NE cells positive for Ki67 (fig. S3, B and D), and very few arose from NE cell replication (4.41% of lineage-labeled NE cells were BrdU<sup>+</sup>) (fig. S3, E and F). Thus, NE cell replication does not account for the accumulation of solitary NE cells.

## Basal stem cells are the source of hypoxia-induced solitary NE cells

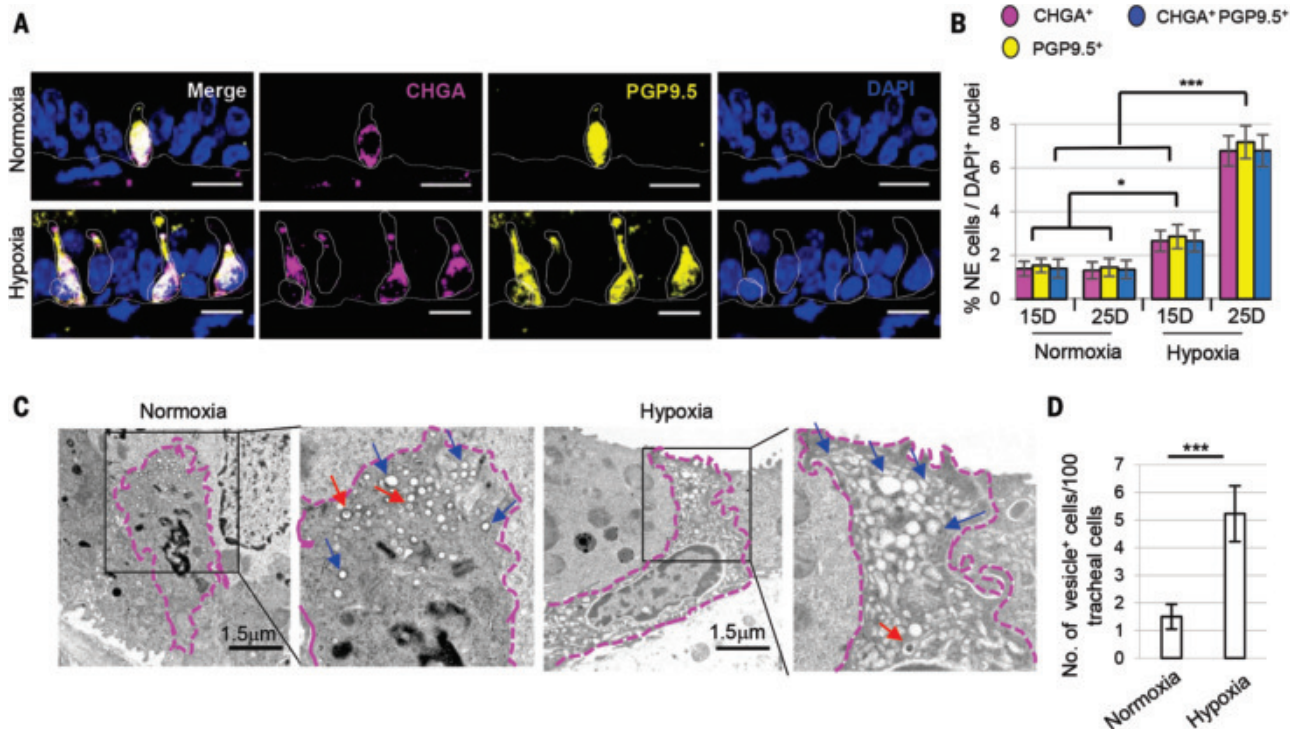
To assess whether basal stem cells are the source of hypoxia-induced solitary NE cells, we lineage labeled basal stem cells using *p63-CreER::R26R-tdTomato* mice (hereafter referred to as *p63-tdTomato*, 97% labeling efficiency) and

<sup>1</sup>Center for Regenerative Medicine, Massachusetts General Hospital, 185 Cambridge Street, Boston, MA 02114, USA.

<sup>2</sup>Departments of Internal Medicine and Pediatrics, Pulmonary and Critical Care Division, Massachusetts General Hospital, Boston, MA 02114, USA. <sup>3</sup>Harvard Stem Cell Institute, Cambridge, MA 02138, USA. <sup>4</sup>Northeastern University, 360 Huntington Ave., Boston, MA 02115, USA. <sup>5</sup>Department of Anesthesia, Critical Care, and Pain Medicine, Massachusetts General Hospital, Boston, MA 02114, USA. <sup>6</sup>Department of Molecular Biology and Howard Hughes Medical Institute, Massachusetts General Hospital, Boston, MA, USA.

<sup>7</sup>Department of Systems Biology, Harvard Medical School, Boston, MA, USA. <sup>8</sup>Broad Institute of Harvard and MIT, Cambridge, MA, USA. <sup>9</sup>Program in Membrane Biology and Division of Nephrology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. <sup>10</sup>Department of Molecular and Cellular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA.

\*Corresponding author. Email: jrajagopal@partners.org



**Fig. 1. Hypoxia leads to an increase in solitary neuroendocrine (NE) cell numbers in the adult mouse trachea.** (A) Immunostaining for NE cell markers PGP9.5 (yellow) and CHGA (magenta). (B) Quantification of the percentage of CHGA<sup>+</sup>, PGP9.5<sup>+</sup> and double-positive solitary NE cells in tracheal sections ( $n = 8$ ). Dotted lines indicate basement membrane and cell borders. (C) Transmission electron microscopy (TEM) images of solitary NE cells and (D) quantification of

cells with NE vesicles after 25 days of normoxia and hypoxia ( $n = 2$ ) (total 300 nuclei from each condition). Blue arrows indicate empty core vesicles, and red arrows point to empty degranulating vesicles. Dotted magenta lines demarcate NE cell borders.  $n =$  biological replicates/condition repeated three times (three independent experiments excluding TEM experiments). \*\*\* $p < 0.001$ , \*\* $p < 0.01$ , \* $p < 0.05$ ; error bars, means  $\pm$  SD. Scale bars, 20  $\mu$ m (unless indicated otherwise).

subjected the mice to hypoxia (fig. S4). This revealed that the basal stem cells are indeed the source of hypoxia-induced solitary NE cells (Fig. 2A). Furthermore, electron microscopy revealed the presence of basal stem cells that expressed abundant vesicles. After hypoxia,  $4 \pm 0.49\%$  of basal cells exhibited NE-specific vesicles (Fig. 2B), which were normally absent. These data point to a direct differentiation of basal cells into solitary NE cells.

Although basal cells clearly contribute to the pool of hypoxia-induced solitary NE cells, a significant increase in the number of unlabeled CHGA<sup>+</sup> tdTomato<sup>+</sup> cells after basal cell lineage tracing was observed, implying contributions from other cells (Fig. 2A). Because airway secretory cells have been shown to be plastic (22), we lineage labeled secretory cells using *Scgb1a1-CreERT2::Rosa26-tdTomato* mice (labeling efficiency  $62 \pm 5.02\%$ ) and then subjected the mice to hypoxia (fig. S5A). A small increase in the fraction of lineage-labeled CHGA<sup>+</sup> solitary NE cells occurred (fig. S5B), revealing that secretory cells can undergo limited NE cell differentiation. Of the non-neuroendocrine cells contributing to the pool of hypoxia-induced solitary NE cells, basal cells are the major contributors (76%), followed by secretory cells (4 to 8%) (fig. S5C).

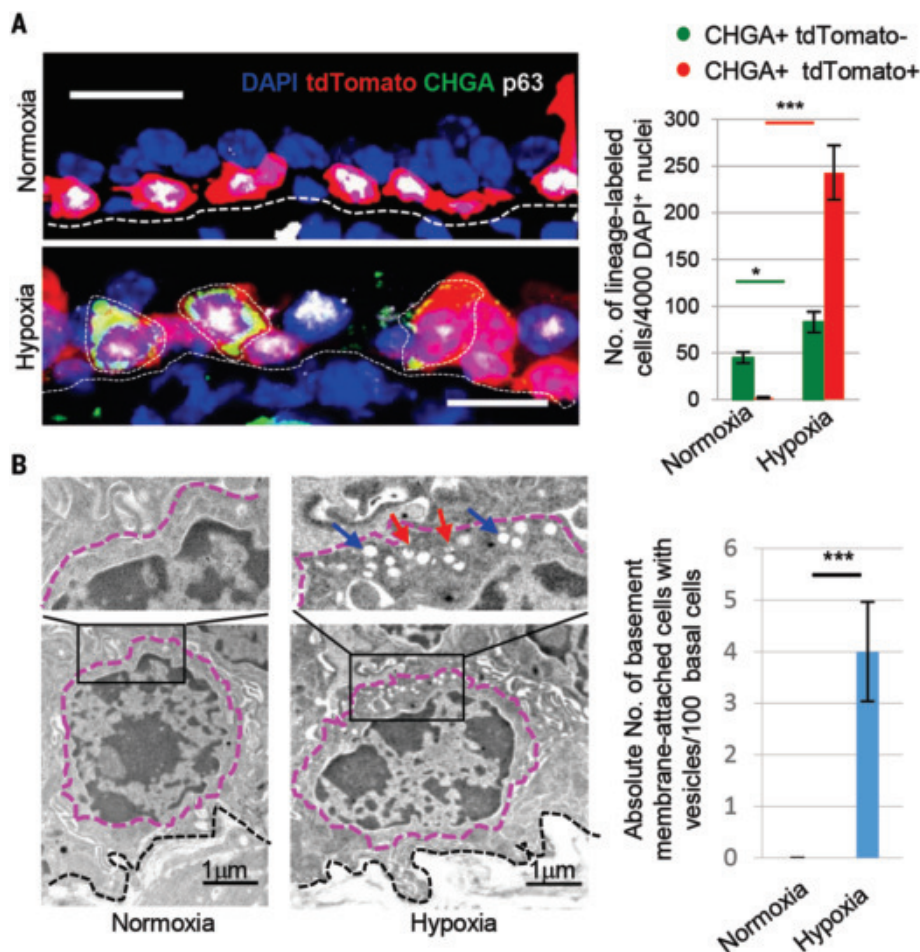
### Basal stem cells sense hypoxia

We next sought to determine the molecular basis of NE cell differentiation. Oxygen-requiring PHDs act as molecular oxygen sensors (23). During normoxia, PHDs hydroxylate hypoxia-inducible factor alpha (HIF $\alpha$ ), which in turn leads to HIF $\alpha$  ubiquitination and degradation. This fails to occur during hypoxia, and stabilized HIF $\alpha$  activates numerous genes associated with the hypoxia response. To address whether PHDs play a similar role in the airway epithelium, we used small molecules that stabilize HIFs in a hypoxia-mimetic state, including the PHD inhibitor FG-4592 and a HIF- $\alpha$ -pVHL interaction inhibitor, CoCl<sub>2</sub> (24, 25). Both inhibitors stabilized HIF-1 $\alpha$  and HIF-2 $\alpha$  in cultured epithelium (fig. S6, A to C). *Ascl1-tdTomato* mice were treated with varying doses of FG-4592 over 20 days, and then tamoxifen was administered to label *Ascl1*<sup>+</sup> solitary NE cells (fig. S7A). A significant dose-dependent increase in tdTomato<sup>+</sup> cells occurred (fig. S7, B and C). Analogous results were obtained with an independent NE cell lineage driver, *Cgrrp-CreERT2::Rosa26-tdTomato* (hereafter *Cgrrp-tdTomato*) (fig. S7, D and E). A significant increase in solitary NE cells was also observed after CoCl<sub>2</sub> administration (fig. S7, F to H).

To assess whether molecular oxygen sensing specifically occurs within the stem cell compartment per se, we genetically activated HIF signaling in vivo by conditionally deleting all three PHDs exclusively in basal stem cells using mice harboring *p63-CreERT2* and *Rosa26-tdTomato::Phd1<sup>fl/fl</sup>::Phd2<sup>fl/fl</sup>::Phd3<sup>fl/fl</sup>* alleles (hereafter referred as *p63-Phds*) (fig. S8). This stabilized HIF-1 $\alpha$  and HIF-2 $\alpha$  in a two-dimensional culture system (fig. S9, A to C). After tamoxifen-induced deletion, CHGA<sup>+</sup> NE cell numbers increased (Fig. 3A). This demonstrates that stem cells harbor a molecular oxygen-sensing mechanism and that stem cell-specific HIF activation triggers those same stem cells to directly differentiate into NE cells.

To identify the specific HIF isoform involved in solitary NE cell differentiation, we specifically activated either *Hif1a* or *Hif2a* by conditionally deleting one *Hif* at a time in basal stem cells (*p63-CreERT2::Rosa26-tdTomato::Hif1a<sup>fl/fl</sup>* or *Hif2a<sup>fl/fl</sup>*) and stabilizing the other through FG-4592 administration (fig. S10). Enforced HIF1 $\alpha$  stabilization and *Hif2a* deletion augmented FG-4592-induced CHGA<sup>+</sup> NE cell induction, whereas HIF2 $\alpha$ -stabilization and *Hif1a* deletion completely blocked it (Fig. 3B). Thus, the stem cell NE





**Fig. 2. Hypoxia leads to increased solitary NE cell numbers by accelerating the differentiation of basal stem cells into NE cells.** (A) Immunostaining for CHGA (green) and p63 (white) and quantification of the absolute number of indicated cells from *p63- tdTomato* lineage-labeled mice exposed to hypoxia. (B) Transmission electron microscopy images of tracheal sections after 25 days of normoxia or hypoxia. After hypoxia, cells with characteristic features of basal cells contain apically located secretory vesicles. Basal cells with and without secretory vesicles are quantified (200 nuclei from each condition). Dotted black lines indicate basement membrane, magenta lines demarcate cell boundaries. Blue arrows point to empty core vesicles, and red arrows point to empty degranulated vesicles. n.s., not significant; \*\*\* $p < 0.001$ , \* $p < 0.05$ ; error bars, means  $\pm$  SD. Scale bars, 20  $\mu$ m (unless indicated otherwise).

differentiation program is *Hif1a* dependent, and *Hif2a* is a negative regulator of this process.

#### Hypoxia-induced solitary NE cells mediate a protective tissue response

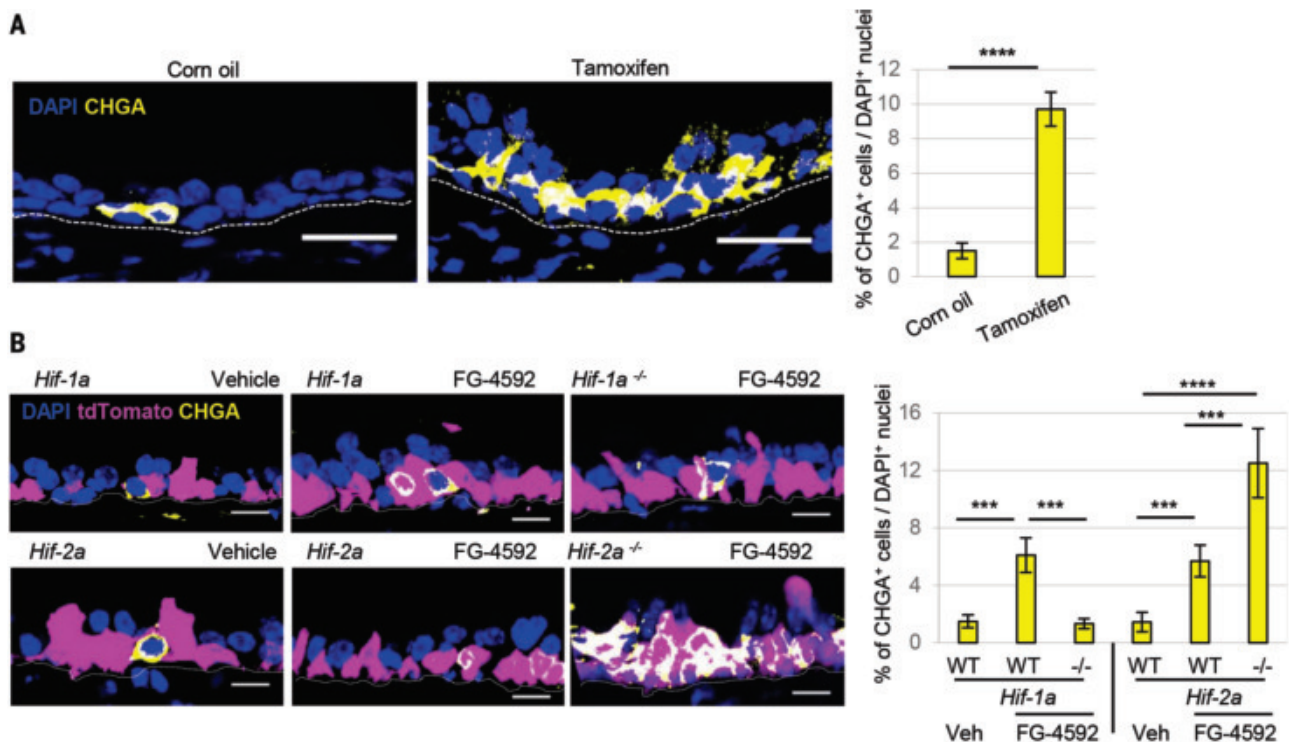
The epithelial damage induced by hypoxia is characterized by epithelial cell apoptosis and a compensatory stem cell hyperplasia (Fig. 4, A and B, columns 1 and 3, and Fig. 5, A and B). We hypothesized that hypoxia-induced solitary NE cells could mediate a protective tissue response in this setting. To assess the functional consequences of solitary NE cell loss, we used compound mice carrying the NE cell-specific *CGRP-CreER* driver allele and a floxed DTA allele. Intranasal administration of a low dose of 4-hydroxytamoxifen was used to achieve airway-specific NE cell ablation, as evidenced by a reduction in CHGA staining (efficiency  $\sim 88 \pm$

2%) (Fig. 5C). NE cell ablation during normoxia did not significantly alter epithelial apoptosis or cell proliferation (Fig. 4B, columns 1 and 2, and Fig. 5, A and B), whereas in the setting of hypoxia, it resulted in a marked increase in apoptosis (Fig. 4B columns 3 and 4, and Fig. 5A) and a drop in epithelial cell proliferation (Fig. 4B, column 3 and 4, and Fig. 5B). Ciliated cells were most prone to apoptosis, whereas basal stem cells were the most hypoxia resistant (fig. S11, A and B). Basal cells were also the dominant cell population undergoing hypoxia-induced proliferation, although secretory cells also divided (fig. S12, A and B). NE cell ablation has no effect during normoxia, but in the setting of hypoxia, the loss of NE cells results in a significant decrease in basal stem cell and secretory cell progenitor cell proliferation (fig. S12, A and B).

Next, we tried to identify an NE cell-associated factor that promotes stem and progenitor cell hyperplasia and/or prevents apoptosis in the setting of hypoxia. We noted that CGRP levels increase after hypoxia (26) and that CGRP serves as a mitogen in rat lung alveolar epithelial cells (27) and murine sub-mucosal gland progenitor cells (28). We then confirmed that epithelial CGRP expression in the trachea is found exclusively in *Ascl1*-positive solitary NE cells (fig. S13, A and B). At baseline, 54% of *Ascl1*-positive NE cells are CGRP immunoreactive, which sharply increases to 92% after hypoxia (fig. S13C). Nasal administration of CGRP to normoxic wild-type mice resulted in increased epithelial cell proliferation (fig. S13D). Costaining with cell type-specific markers and Ki67 identified replicating basal stem cells ( $CK5^+ Ki67^+$ ) and a smaller increase in replicating secretory cells ( $SCGB1A1^+ Ki67^+$ ) (fig. S13, D to F). Furthermore, we were able to block CGRP-stimulated proliferation using the CGRP-receptor inhibitor BIBN-4096 (29) (fig. S13, D to F).

Next, we intranasally administered CGRP peptide after NE cell ablation to assess whether the presence of CGRP could compensate for the loss of whole NE cells. CGRP restored epithelial proliferation to amounts normally observed after hypoxia and also prevented epithelial apoptosis (Fig. 4B, columns 3 to 5, and Fig. 5, A and B). Furthermore, proliferation was restored in both basal stem cells and secretory cells (fig. S14, A and B). Ciliated cells were the predominant cell type protected from apoptosis, although basal stem cells were also protected (fig. S11B). The CGRP-receptor inhibitor BIBN-4096 appropriately blocked the CGRP-mediated effects on cell type-specific apoptosis and proliferation (Fig. 4B, columns 5 and 6, and figs. S11A, columns 5 and 6, and S14). TUNEL (terminal deoxynucleotidyl transferase-mediated deoxyuridine triphosphate nick end labeling) staining confirmed these findings (fig. S15, A and B).

To determine whether CGRP sourced from cells other than solitary NE cells contributes to epithelial protection, we assessed which other cell types were ablated in our *Cgrp-CreERT2::Rosa26-DTA* model. We noted that tracheal innervation was lost, so we used an *Ascl1-CreERT2::Rosa26-DTA* model to eliminate solitary NE cells and spare CGRP-expressing neuronal populations (fig. S16). The results closely mirrored our findings with the *Cgrp-CreERT2::Rosa26-DTA* model (fig. S17). Additionally, the administration of CGRP inhibitor to hypoxic animals with intact solitary NE cells resulted in a significant increase in apoptosis and a drop in proliferation, consistent with a protective role for CGRP (fig. S17, B, C, columns 5 and 7, D, and E). Moreover, hypoxia led to increased epithelial expression of the CGRP receptors RAPM1 and CALCRL (fig. S18, A and B), suggesting increased CGRP signal reception.



**Fig. 3. Stem cells sense hypoxia using prolyl hydroxylases and then differentiate into solitary NE cells. (A)** Immunostaining for CHGA (yellow) and quantification of NE cells after deletion using *p63-CreER::PhDs* mice. **(B)** Immunostaining for CHGA (yellow) and quantification of CHGA+ cells from

*p63-Hif1α* or *p63-Hif2α* mice treated with FG-4592 (*n* = 5). In the merged image, the overlap of yellow (CHGA), blue (DAPI, 4',6-diamidino-2-phenylindole), and tdTomato appears white. \*\*\*\**p* < 0.0001, \*\*\**p* < 0.001; error bars, means ± standard deviation. Scale bars, 20 μm.

We next sought to assess whether CGRP secreted by NEBs was acting as a long-range protective factor, originating in the intrapulmonary airway epithelium where NEBs occur, and acting directly or indirectly on the remote tracheal epithelium. Using *Ascl1-tdTomato* mice to lineage label preexisting NEBs, we confirmed that hypoxia triggers NEB hyperplasia, consistent with the previously reported finding that NEBs lacking PHDs become larger (20). Although there was no significant increase in the total number of NEBs (fig. S19, A to D), scattered NEBs did contain rare Ki67 cells (4 to 10 NEBs per left lobe of the lung) (fig. S19C). However, despite the efficient ablation of tracheal solitary NE cells in our model system, NEBs were unaffected (fig. S20). This suggests that solitary NE cells are the source of protective CGRP in our model system and that this peptide acts locally to protect adjacent epithelial cells from hypoxia-induced damage. However, these findings do not exclude a physiologic role for pools of CGRP produced elsewhere.

Finally, we assessed the effects of hypoxia on human airway epithelia grown from airway basal stem cells. Stabilizing HIFs by adding FG-4592 significantly increased the number of solitary NE cells (fig. S21). This system should prove useful in defining the role of solitary NE cells in diseases characterized by NE cell excess, paralleling the use of such models

in dissecting the root mechanisms of cystic fibrosis.

### Discussion

We have shown that hypoxia-induced murine solitary NE cells are necessary for repairing hypoxia-induced epithelial damage. The mechanism invokes the secretion of a protective paracrine signal. We speculate that the diffuse distribution of solitary NE cells throughout the human airway, in contrast to the discrete location of murine NEBs at airway bifurcations, forms the basis of a distributed tissue-wide protection system in which effective epithelial repair can be fostered throughout the length of the airway tree.

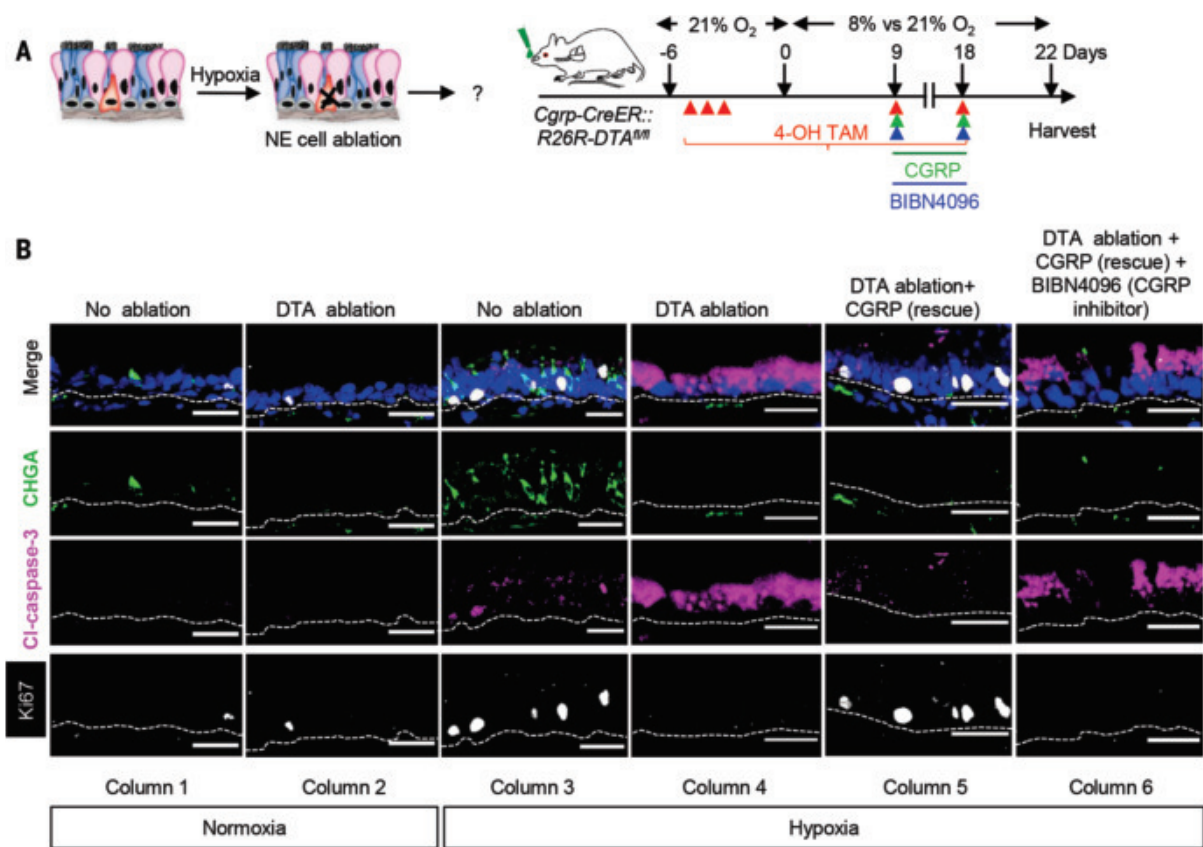
Oxygen therapy is used in the setting of respiratory failure in many diseases associated with NE cell excess, ranging from severe asthma to cystic fibrosis to COPD, but oxygen has also been associated with multiple forms of toxicity (30). If hypoxia-induced solitary NE cells are indeed protective in disease states, supportive oxygen therapy might result in unintended consequences by reducing the physiologic stimulus for generating protective NE cells. By contrast, in neuroendocrine hyperplasia of infancy (NEHI), where NE cell excess appears to be a primary pathology, supplemental oxygen might act as a primary therapy by triggering a reduction in pathologic NE cell numbers. Indeed, some

patients with NEHI have a sudden improvement with oxygen administration that is uncharacteristic of related respiratory disorders of childhood, which display a slow and graded response to oxygen therapy (31).

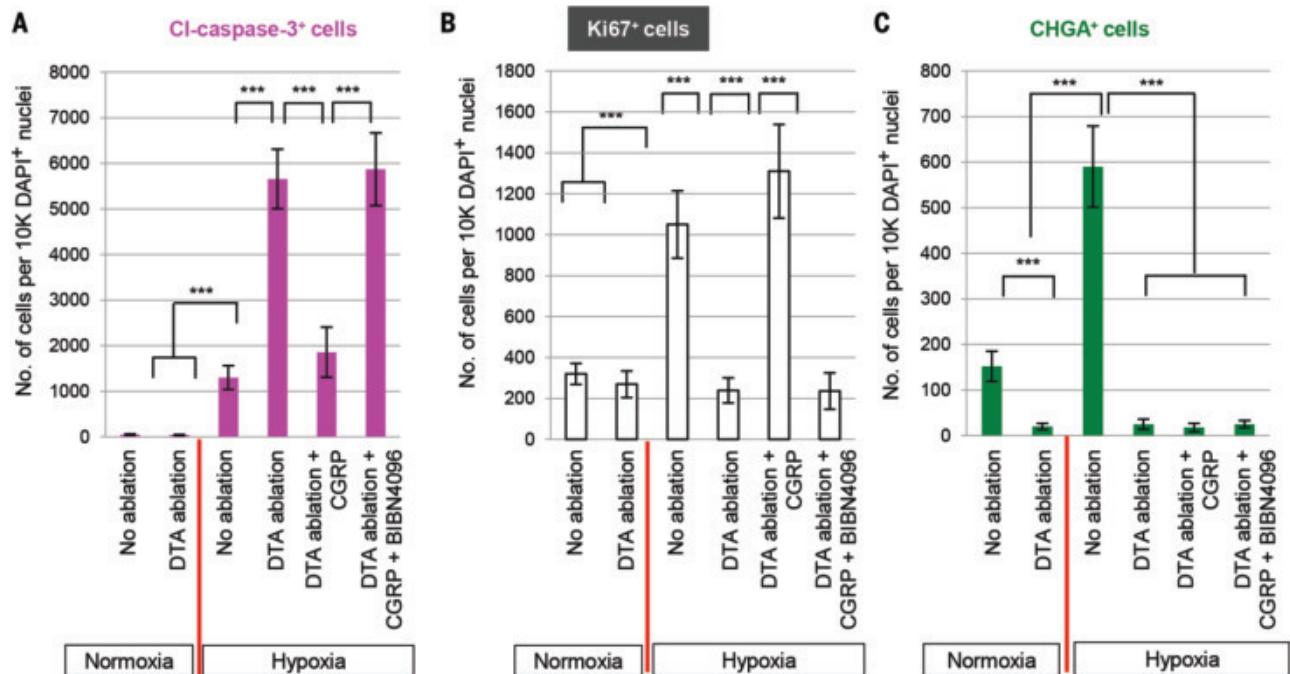
With regards to normal physiology, our findings may help explain why NE cell numbers decrease after birth (17–19, 32) because oxygen levels increase after air breathing, as well as why high altitude is associated with increased numbers of NE cells (33). It may also explain why Notch loss-of-function experiments produce increased numbers of NE cells in the embryonic epithelium, whereas they do not in the air-breathing adult. Finally, it is of interest to note that sudden infant death syndrome (SIDS) has been associated with hypoxia, NE cell excess, and elevated serotonin levels (34, 35).

In this study, we have identified a conserved tissue-level response to a fundamental form of stress, oxygen deprivation. In this instance, sentinel stem cells detect hypoxia and produce a specific protective cell type that is needed to mitigate the effects of the hypoxia. Although we have identified CGRP as a protective airway solitary NE cell factor, NE cells can secrete a host of other neuropeptides and amines. Thus, we speculate that different forms of pulmonary injury might engender varied protective NE cell responses. It will be of great interest to assess whether other organ-specific stem cells





**Fig. 4. Solitary NE cells are required for the repair of hypoxia-induced injury.** (A) Schematic representation of the protocol for assessing the functional consequences of NE cell ablation in the setting of hypoxia-induced injury. (B) Immunostaining for CHGA (green), Cleaved caspase-3 (magenta) and Ki67 (white) on normoxic and hypoxic tracheal epithelium with and without NE cell ablation and in combination with CGRP rescue and CGRP receptor inhibitor administration. ( $n = 5$ ). \*\*\* $p < 0.001$ , \*\* $p < 0.01$ , \* $p < 0.05$ ; error bars, means  $\pm$  standard deviation. Scale bars 30  $\mu$ m.



**Fig. 5. CGRP stimulates progenitor replication and prevents epithelial apoptosis.** Quantification of (A) Cleaved caspase-3+ cells, (B) Ki67+, and (C) CHGA<sup>+</sup> cells under normoxic or hypoxic conditions with and without NE cell ablation coupled to CGRP administration or blockade. Tissue samples from Fig. 4. \*\*\* $p < 0.001$ ; error bars, means  $\pm$  SD.

more generally execute their own specific protective behaviors when triggered by hypoxia. It will also be important to determine whether the stem cells of other organs respond to local hypoxia by generating their own distinct populations of protective NE cells.

## REFERENCES AND NOTES

1. F. Feyrter, *Virchows Arch. Pathol. Anat. Physiol.* **320**, 551–563 (1951).
2. F. Feyrter, *Virchows Arch. Pathol. Anat. Physiol. Med.* **325**, 723–732 (1954).
3. E. Cutz, *Semin. Diagn. Pathol.* **32**, 420–437 (2015).
4. E. Cutz, H. Yeger, J. Pan, T. Ito, *Curr. Respir. Med. Rev.* **4**, 174–186 (2008).
5. M. Noguchi, K. Sumiyama, M. Morimoto, *Cell Rep.* **13**, 2679–2686 (2015).
6. C. S. Kuo, M. A. Krasnow, *Cell* **163**, 394–405 (2015).
7. E. Cutz, A. Jackson, *Respir. Physiol.* **115**, 201–214 (1999).
8. R. I. Linnoila, *Lab. Invest.* **86**, 425–444 (2006).
9. E. Cutz, J. Pan, H. Yeger, N. J. Domnik, J. T. Fisher, *Semin. Cell Dev. Biol.* **24**, 40–50 (2013).
10. K. Branchfield et al., *Science* **351**, 707–710 (2016).
11. S. D. Reynolds et al., *Am. J. Physiol. Lung Cell. Mol. Physiol.* **278**, L1256–L1263 (2000).
12. X. Gu et al., *Am. J. Respir. Cell Mol. Biol.* **50**, 637–646 (2014).
13. Y. Ouadah et al., *Cell* **179**, 403–416.e23 (2019).
14. H. Song et al., *Proc. Natl. Acad. Sci. U.S.A.* **109**, 17531–17536 (2012).
15. D. T. Montoro et al., *Nature* **560**, 319–324 (2018).
16. J. K. Watson et al., *Cell Rep.* **12**, 90–101 (2015).
17. I. M. Keith, J. A. Will, *Thorax* **36**, 767–773 (1981).
18. N. S. Track, E. Cutz, *Life Sci.* **30**, 1553–1556 (1982).
19. H. Moosavi, P. Smith, D. Heath, *Thorax* **28**, 729–741 (1973).
20. J. Pan, T. Bishop, P. J. Ratcliffe, H. Yeger, E. Cutz, *Hypoxia (Auckl.)* **4**, 69–80 (2016).
21. J. R. Rock, S. H. Randell, B. L. M. Hogan, *Dis. Model. Mech.* **3**, 545–556 (2010).
22. P. R. Tata et al., *Nature* **503**, 218–223 (2013).
23. C. Willam, L. G. Nicholls, P. J. Ratcliffe, C. W. Pugh, P. H. Maxwell, *Adv. Enzyme Regul.* **44**, 75–92 (2004).
24. K. Wu et al., *Brain Res.* **1632**, 19–26 (2016).
25. Z.-J. Dai et al., *J. Exp. Clin. Cancer Res.* **31**, 28 (2012).
26. D. R. Springall et al., *J. Pathol.* **155**, 259–267 (1988).
27. Y. Kawanami et al., *Respir. Res.* **10**, 8 (2009).
28. W. Xie et al., *J. Clin. Invest.* **121**, 3144–3158 (2011).
29. H. Doods et al., *Br. J. Pharmacol.* **129**, 420–423 (2000).
30. P. Tinitis, *Ann. Emerg. Med.* **12**, 321–328 (1983).
31. S. Caimmi et al., *Ital. J. Pediatr.* **42**, 84 (2016).
32. E. R. Spindel et al., *J. Clin. Invest.* **80**, 1172–1179 (1987).
33. J. R. Gosney, *Anat. Rec.* **236**, 105–107, discussion 108–112 (1993).
34. K. Aita et al., *Leg. Med.* **2**, 134–142 (2000).
35. P. M. A. Siren, *Ups. J. Med. Sci.* **121**, 199–201 (2016).

## ACKNOWLEDGMENTS

We thank the members of the Rajagopal lab and R. Chivukula for constructive criticism. We thank P. Chuang for providing *Cgrp-CreER*. We thank New England Donor Services (NEDS) for providing human airway samples. Finally, we thank E. Cutz for his constructive criticism and for reminding us of the history of NE cell biology. **Funding:** J.R. is supported by the New York Stem Cell Foundation, the National Institutes of Health–National Heart, Lung, and Blood Institute (R01HL118185, and R01HL148351-01A1), and the Ludwig Cancer Institute at Harvard. J.X. is supported by R01CA193455. Electron microscopy was performed in the Microscopy Core of the MGH Program in Membrane Biology, which is partially supported by an Inflammatory Bowel Disease Grant (DK043351) and a Boston Area Diabetes and Endocrinology Research Center (BADERC) Award (DK057521). **Author contributions:** M.S. conceived, designed, and performed the experiments and co-wrote the manuscript; U.K.C. performed immunostaining and counting cell numbers; R.M.H.G., I.H.J., and F.I. provided guidance on the use of hypoxia chambers; D.C. performed electron microscopy imaging; L.L. and J.X. provided *p63-CreER* mice; W.M.Z. and V.K.M. provided guidance on hypoxia experiments; and J.R. supervised the work and co-wrote the manuscript. All authors reviewed and edited the manuscript for accuracy. **Competing interests:** V.K.M. is a paid adviser to Janssen Pharmaceuticals and 5am Ventures and is a founder of and owns equity in Raze Therapeutics. V.K.M., I.H.J., and W.M.Z. are listed as inventors on patent application (W02017027810A2) submitted by Massachusetts General Hospital on therapeutic uses of hypoxia for mitochondrial disease. **Data and materials availability:** *p63-CreER* mice are available from J. Xu under

a materials transfer agreement with the Baylor College of Medicine. All data are available in the manuscript or the supplementary materials.

## SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/371/6524/52/suppl/DC1  
Materials and Methods

Figs. S1 to S21  
References (36–40)  
MDAR Reproducibility Checklist

31 October 2019; accepted 29 October 2020  
10.1126/science.aba0629

## PROTEIN SYNTHESIS

# Interactions between nascent proteins translated by adjacent ribosomes drive homomer assembly

Matilde Bertolini<sup>1\*</sup>, Kai Fenzl<sup>1\*</sup>, Ilia Kats<sup>1†</sup>, Florian Wruck<sup>2</sup>, Frank Tippmann<sup>1</sup>, Jaro Schmitt<sup>1</sup>, Josef Johannes Auburger<sup>1</sup>, Sander Tans<sup>2,3</sup>, Bernd Bukau<sup>4,†</sup>, Günter Kramer<sup>1,†</sup>

Accurate assembly of newly synthesized proteins into functional oligomers is crucial for cell activity. In this study, we investigated whether direct interaction of two nascent proteins, emerging from nearby ribosomes (co-co assembly), constitutes a general mechanism for oligomer formation. We used proteome-wide screening to detect nascent chain–connected ribosome pairs and identified hundreds of homomer subunits that co-co assemble in human cells. Interactions are mediated by five major domain classes, among which N-terminal coiled coils are the most prevalent. We were able to reconstitute co-co assembly of nuclear lamin in *Escherichia coli*, demonstrating that dimer formation is independent of dedicated assembly machineries. Co-co assembly may thus represent an efficient way to limit protein aggregation risks posed by diffusion-driven assembly routes and ensure isoform-specific homomer formation.

Sophisticated mechanisms have evolved to ensure efficient and accurate protein complex biogenesis, including the fine-tuning of subunit expression to match complex stoichiometries (1), the employment of general or dedicated chaperones to guide oligomerization (2–4), the colocalization of subunit synthesis (5–7), and the timely oligomerization by coupling translation and subunit interactions (cotranslational assembly) (3, 8, 9). Selective ribosome profiling (SeRP) has provided mechanistic details of cotranslational assembly for *Vibrio harveyi* luciferase expressed in *Escherichia coli* (3) and several heteromeric complexes in yeast (8). In all cases studied, a freely diffusing, presumably folded protein engages its nascent partner subunit (co-post assembly).

In this study, we tested whether cotranslational assembly of protein complexes may also occur via association of two nascent subunits concurrently translated by two ribosomes (co-co assembly). A priori, co-co assembly may involve nascent chains synthesized on two different mRNAs (in trans) or, for homo-oligomer assembly, on the same mRNA (in cis). Notably, cis

assembly does not require that distinct mRNA molecules colocalize in the cytosol and enables transcript-specific homomeric complex generation, avoiding undesired interactions between closely related proteins or wild-type and mutant alleles (10). Although co-co assembly has already been proposed for individual protein complexes in different organisms (10–14), direct experimental evidence that two ribosome–nascent chain complexes interact is still missing, and we lack any information on the prevalence, molecular mechanisms, and relevance of this proposed assembly process. We thus developed disome selective profiling (DiSP)—an unbiased, proteome-wide screening based on ribosome profiling (15)—to reveal the co-co assembly proteome in human cells.

## DiSP reveals widespread disome formation mediated by nascent chain interactions

To identify co-co assembling complexes across the proteome, we reasoned that ribosome pairs (disomes) connected by their exposed nascent chains will remain connected even upon mRNA digestion. Thus, it should be possible to detect co-co assembly candidates by ribonuclease (RNase) treatment of cell lysates, followed by separation of monosomes and disomes in sucrose gradients and deep sequencing of 30-nucleotide (nt) ribosomal footprints from both fractions (DiSP; Fig. 1A and fig. S1A). The disome fraction will also contain RNase-resistant disomes that form upon collision of ribosomes that translate the same mRNA; however, these disomes will protect double-length (60-nt) mRNA fragments (16) and are not analyzed by

<sup>1</sup>Center for Molecular Biology of Heidelberg University (ZMBH) and German Cancer Research Center (DKFZ), DKFZ-ZMBH Alliance, Im Neuenheimer Feld 282, Heidelberg D-69120, Germany. <sup>2</sup>AMOLF, Science Park 104, 1098 XG Amsterdam, Netherlands. <sup>3</sup>Department of Bionanoscience, Delft University of Technology and Kavli Institute of Nanoscience Delft, 2629HZ Delft, Netherlands.

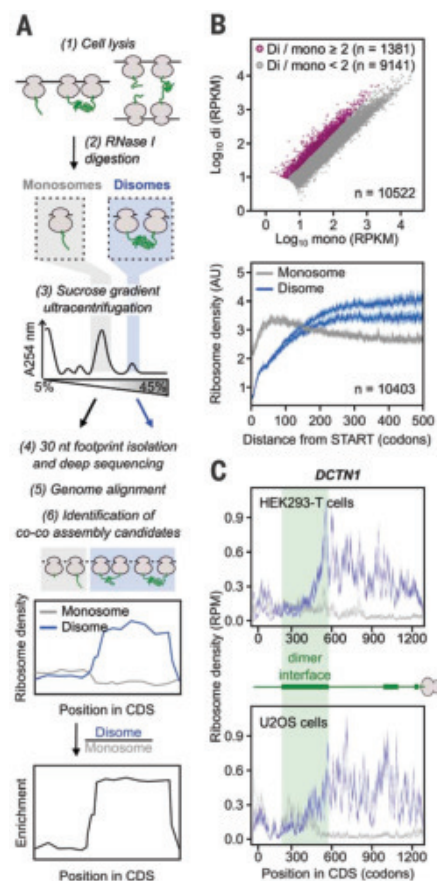
\*These authors contributed equally to this work.

†Present address: Computational Genomics and System Genetics, DKFZ, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany.

‡Corresponding author. Email: g.kramer@zmbh.uni-heidelberg.de (G.K.); bukau@zmbh.uni-heidelberg.de (B.B.)



DiSP. Translating ribosomes engaged in co-co assembly will shift from the monosome to the disome fraction upon nascent chain dimerization, which could be detected by analyzing the relative footprint density of both sam-



**Fig. 1. Disome selective profiling (DiSP) reveals widespread disome formation.** (A) Experimental procedure of DiSP. Cell lysates are treated with RNase I (1 and 2); monosomes and disomes are separated by sucrose gradient ultracentrifugation (3); and ~30-nt-long ribosome footprints are extracted, converted into a DNA library, and sequenced (4). Co-co assembly candidates are identified by a shift of the footprint density from monosome to disome fraction, or by a disome-over-monomosome enrichment profile (5 and 6). A254 nm, absorbance at 254 nm. (B) Comparison of disome (di) and monosome (mono) footprint density of all detected genes in HEK293-T cells (top; one replicate shown). Average footprint density along the coding sequence of all detected genes (metagene) aligned to the start of translation (bottom; two biological replicates). RPKM, reads per kilobase per million mapped reads. (C) Monosome (gray) and disome (blue) footprint density along the coding sequence (CDS) of *DCTN1*. The cartoon shows exposed nascent chain segments during translation; green bars indicate dimerization interfaces. DiSP data of HEK293-T cells (two biological replicates) and U2OS cells (two biological replicates) are compared. RPM, reads per million.

ples (separately or as enrichment of disome over monosome) along a gene's coding sequence (Fig. 1A). In contrast to SeRP, which has been used to explore co-post assembly of selected protein complexes (3, 8), DiSP can provide proteome-wide interaction profiles of all translating ribosomes.

We initially performed DiSP of human embryonic kidney 293-T (HEK293-T) cells. To identify co-co assembly candidates, we first compared gene-specific footprint densities in the disome and monosome fractions, revealing more than 1300 genes with a disome-over-monomosome enrichment value  $\geq 2$  (Fig. 1B, top). A metagene profile of the averaged monosome and disome density along all coding sequences showed that early during translation, when nascent chains are short, ribosomes mostly migrated as monosomes, followed by a steady disome enrichment that leveled out at ~200 codons (Fig. 1B, bottom). The monosome-to-disome shift of translating ribosomes occurred only in a subset of genes, supporting the assumption that it depended on interaction properties of nascent chains (Fig. 1B, top, and fig. S1B). One example among the twofold disome enriched genes is *DCTN1*, which encodes p150<sup>glued</sup>, a subunit of the dyactin motor complex. Ribosomes that translate *DCTN1* convert from monosomes to disomes near codon 430, when ~400 amino acids of nascent p150<sup>glued</sup> are exposed on the ribosomal surface. This N-terminal segment includes major parts of the coiled-coil dimerization domain, suggesting that the disome shift was caused by cotranslational homodimerization (Fig. 1C, top). Repeating DiSP in U2OS cells, we found a large overlap of disome-enriched genes and robustly correlated enrichment profiles (Fig. 1C and fig. S1, B and C), demonstrating that disome formation is a general feature of a specific subset of nascent proteins across different cell types.

To challenge our model that disome formation is mediated by nascent proteins, we explored whether disome shifts were sensitive to release or degradation of nascent chains. Treatment of lysates with puromycin (Puro) or increasing concentrations of proteinase K (PK) efficiently suppressed the shift of footprints from monosome to disome. This was apparent from a general reduction of the disome enrichment (Fig. 2A) and a flattening of enrichment profiles at the metagene level (Fig. 2B) and for individual genes (Fig. 2C and fig. S1, D to G). Thus, the stability of DiSP-detected disomes critically depends on the integrity of nascent chains, in agreement with the model of co-co assembly.

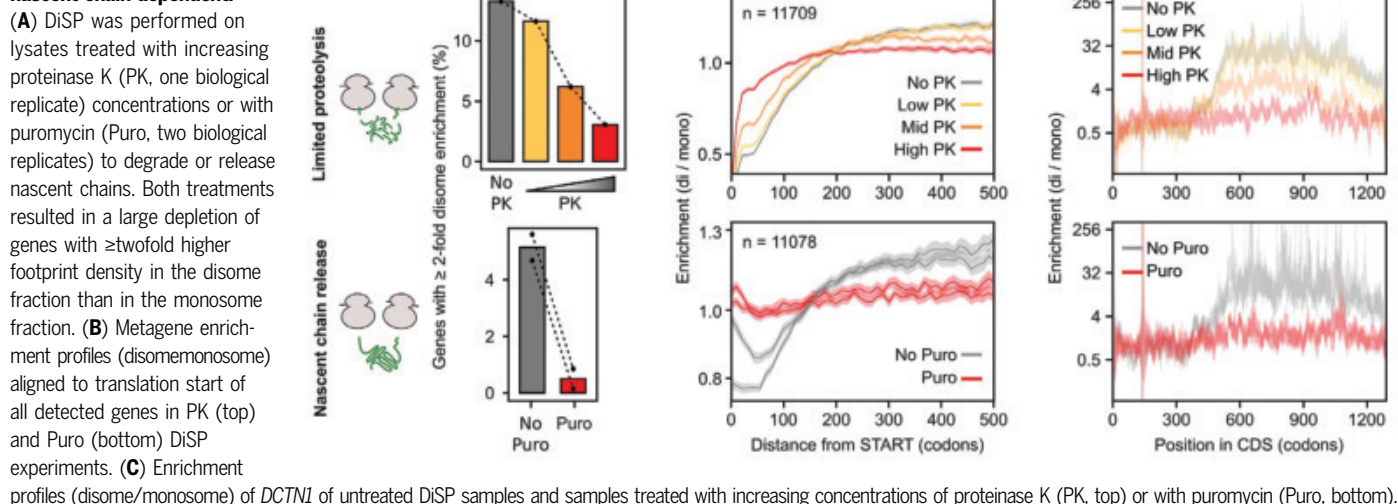
#### A high-confidence list of co-co assembly candidates enriched for homomers

We developed an unbiased bioinformatics selection regime to classify proteins on the

basis of their proficiency to co-co assemble. Accordingly, a protein qualified as a high-confidence candidate if all of the following criteria were fulfilled: (i) The gene's enrichment profile had a sigmoidal shape, indicating that with progressing translation, ribosomes shifted from the monosome to the disome fraction. If one of the interacting ribosomes terminates earlier, the other ribosome in the pair will shift back to the monosome fraction before it reaches the end of the coding sequence, resulting in a double-sigmoidal shift (Fig. 3A). (ii) The enrichment profile becomes less sigmoidal upon treatment of the lysate with puromycin and (iii) similarly with PK. (iv) The mature protein localizes to either the cytoplasm or the nucleus. We decided to categorize translocated proteins as low-confidence candidates because we cannot formally exclude the possibility that these ribosomes interact with membrane components of the translocation machinery and therefore migrate in the disome fraction. In addition, our validation experiments focused on cytosolic and nuclear candidates (fig. S4), and poor structural annotation of membrane proteins complicates the downstream bioinformatics analysis. Out of 15,898 detected genes, 829 fulfilled all criteria and were classified as high-confidence co-co assembly candidates (table S1). A large number of genes (3301) fulfilled the important criterion (i) but not all of criteria [(ii) to (iv)] and were therefore categorized as low-confidence candidates (table S1). The low-confidence list included 1404 proteins that are translocated across or inserted into organelle membranes [mainly the endoplasmic reticulum (ER)]; of these, 443 fulfilled all other criteria. The latter fraction reflects the general frequency of ER-translocated proteins in the human proteome and indicates that co-co assembly may be an equally important mechanism for assembly of cytosolic or nuclear and ER complexes, in agreement with previous experimental indications (17–19). The disome shift of ribosomes that synthesize membrane proteins frequently occurs after exposure of the first transmembrane domain (TMD) (fig. S2A), which may suggest that co-co assembly involves interactions of two TMDs in the ER membrane.

Our next aim was to quantitatively assess what fraction of each high-confidence candidate assembles cotranslationally (hereafter termed “efficiency” of co-co assembly). The efficiency was estimated by determining the reduction of footprints in the monosome fraction after initiation of co-co assembly relative to those in the total translatoome [including all translating ribosomes, determined by classical ribosome profiling (15, 20)]. Metagene analyses of footprint densities of all high-confidence genes aligned to the onset of assembly revealed a reduction of footprints in the monosome fraction from a DiSP experiment but not in the

**Fig. 2. Disome formation is nascent chain dependent.**



total translome (Fig. 3B, top). This result confirmed that the monosome depletion was caused by a shift of ribosomes to the disome fraction. The median monosome footprint reduction after the detected co-co assembly onset of high-confidence genes was  $\sim 40\%$ , and for some genes even exceeded  $90\%$ , indicating that, in many cases, most nascent chains assembled cotranslationally (Fig. 3B, bottom). Monosome depletion was also observed (to a smaller extent) for many low-confidence candidates, suggesting that this list includes additional proteins that employ co-co assembly as a main route for complex formation (fig. S2, B and C). Notably, the calculated depletion value most likely underestimates the *in vivo* co-co assembly efficiency because of (i) the inevitable slight cross-contamination between the monosome and disome fractions and (ii) the possibility of a partial loss of disomes, which are connected by comparably weak nascent chain interactions, during sucrose gradient centrifugation. Supporting this notion, the three proteins with the highest efficiency ( $\geq 90\%$  depletion; namely, TPR, EEA1, and CLIP1) contained extremely long coiled-coil homodimerization domains (between 1000 and 1500 amino acids, compared with a median coiled-coil length of 66 amino acids in the cellular proteome), suggesting high stability.

We went on to analyze the features of proteins included in the high- and low-confidence lists. Consistently, annotated monomeric proteins were depleted in both lists of co-co assembly candidates, most extensively among the high-confidence proteins (Fig. 3C and table S2). Both classes showed a significant enrichment of homomers, but heteromers were not significantly enriched. As our statistical analysis accounts for differences in expression levels in our datasets and the annotation database, the heteromer enrichment in the low-confidence class is statistically not significant, although it

slightly exceeds the homomer enrichment. Furthermore, we often found only one subunit of a heterodimer in our candidate list, which suggests that this subunit formed a homo-oligomer or co-co assembled with an as-yet-unknown partner subunit.

We used available crystal structures of protein complexes to determine the position of residues involved in subunit interaction at the onset of the disome shift. This analysis showed that the onset of assembly often coincided with the emergence of nascent chain segments that form the interfaces for the homo-oligomers (Fig. 3D, left). This correlation was not detected for heteromeric high-confidence candidates (Fig. 3D, right). Although these findings do not exclude the possibility that individual heteromers co-co assemble, as previously reported (13, 14, 19), they suggest that co-co assembly is predominantly employed for the formation of homomeric protein complexes.

#### Co-co assembly is driven by exposure of conserved N-terminal homodimerization domains

Most detected co-co assembly interactions were established at early translation stages (fig. S3A). Consistently, homodimerization interfaces are enriched in the N-terminal halves of high-confidence candidates (fig. S3B, left). This is different in the majority of the human proteome, where homodimerization interfaces are more often located in the C-terminal half of the protein, as previously reported (21) (fig. S3B, right).

We next aimed to identify protein motifs or folds that mediate co-co assembly, by studying the enrichment of exposed domains at the onset of assembly. This analysis identified seven domain clusters mediating co-co assembly (color coded in Fig. 4A), of which five are established homodimerization units.

Among our high-confidence candidates, coiled coils were the most prevalent annotated do-

main class that is exposed on the ribosome surface at assembly onset (193 of 829 proteins according to UniprotKB; Fig. 4B, left). Furthermore, the DeepCoil prediction tool (22) identified coiled-coil segments on the exposed nascent chains in 408 genes (fig. S3C), suggesting that up to 50% of high-confidence candidates employ this fold for co-co assembly. In many cases, the coiled coil is only partially exposed at assembly onset (Fig. 4B, left). The number of exposed residues involved in coiled-coil formation varied (median of 111 residues in the high-confidence class; fig. S3D), which may indicate that different lengths of the coiled coil are needed to form a stable dimer.

We found seven additional domains that were generally positioned N-terminally to coiled-coil domains in myosins, kinesins, and AGC kinases (orange in Fig. 4A) and were therefore exposed at the onset of co-co assembly. However, disome enrichment generally required the partial or complete exposure of the coiled-coil segment, suggesting that these domains do not contribute to oligomerization.

A second domain class that was often only partially exposed at the onset of assembly is BAR domains (named after Bin, amphiphysin, and Rvs proteins; Fig. 4B, right). These conserved dimerization domains are found in many proteins that mediate membrane curvature. They consist of three (classical BAR) to five (F-BAR) bent antiparallel  $\alpha$  helices. According to our dataset, co-co assembly generally required the exposure of the most N-terminal  $\alpha$  helix (helix1; Fig. 4B, right), which interacts with its partner (helix1') in an antiparallel fashion.

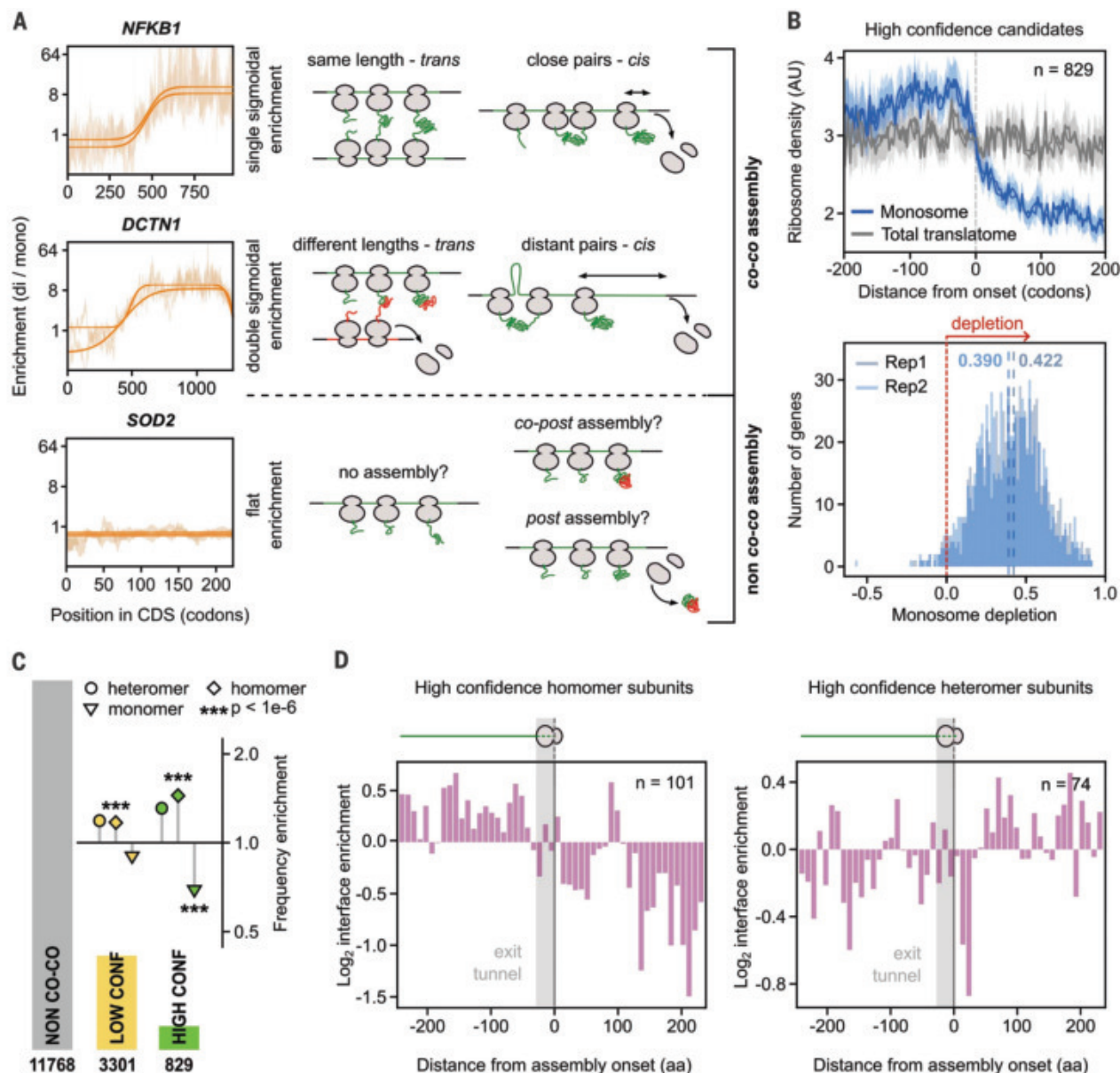
All other enriched domain classes—including BTB (Broad-Complex, Tramtrack, and Bric a brac), RHD (Rel homology domain), and SCAN (SRE-ZBP, CTfin51, AW-1, and Number 18 cDNA) domains (Fig. 4C)—were globular and fully exposed at assembly onset, implying that their



cotranslational folding was required for assembly. BTBs are highly conserved globular dimerization domains located at the N termini of many transcription factors, ion channels, and E3 ligase subunits, and were found in 36 of

our high-confidence candidates (Fig. 4C, left). The less abundant RHDs are found at the N terminus of proteins involved in nuclear factor  $\kappa$ B (NF- $\kappa$ B) complex formation and create the interface of homo- and heteromeric interactions.

According to our DiSP, all NF- $\kappa$ B homologs co-co assemble, confirming earlier indications that proteins encoded by *NFKB1* may cotranslationally assemble in cis and that early assembly is required for native biogenesis of the p50



**Fig. 3. High-confidence co-co assembly proteins are enriched in homo-oligomers.** (A) Examples of gene-specific disome-over-monosome enrichment profiles (DiSP data, in the background; two biological replicates) and the corresponding fitting (solid lines) for each of the three possible shapes of DiSP enrichments. The single sigmoid is consistent with nascent chain-connected ribosomes that terminate translation simultaneously, either by co-co assembly in trans (if the mRNA segments translated by both ribosomes after co-co assembly have similar lengths) or in cis (with ribosomes that closely follow each other on the same mRNA) (top). The double sigmoid is consistent with co-co assembly involving two ribosomes that do not terminate at the same time; this may occur in trans (if the mRNA segments translated by both ribosomes after co-co assembly have different lengths) or in cis (if the leading ribosome is distant from the trailing one) (middle). Flat enrichment profiles indicate that nascent proteins do not co-co assemble. (B) (Top) Metagenes profiles of all high-confidence candidates

aligned to assembly onset. Footprint density in the monosome fraction and the total translome are shown (two biological replicates). (Bottom) Gene-specific quantification of the efficiency of co-co assembly, calculated as the relative depletion of footprint density in the monosome relative to the total translome after assembly onset. The median monosome depletion for each replicate is indicated by blue dashed lines. AU, arbitrary units. (C) Frequency enrichment of annotated subunits of protein complexes in high- and low-confidence lists relative to the whole proteome (absolute and relative numbers are provided in table S2) (31). The number of genes included in each assembly class is indicated in the bar plot. *P* values were calculated using an enrichment test adjusted for expression bias (31, 32). (D) Distribution of residues forming the intersubunit interface of protein complexes determined from available crystal structures. The position of interface residues on the proteins' primary sequence is aligned to assembly onset of high-confidence homomers (left) or heteromers (right). aa, amino acids.

transcription factor (12, 23) (Fig. 4C, middle, and fig. S1B, right). This notion very likely also holds true for the *RELB*-encoded homolog; however, because *RELB* is poorly expressed in HEK293-T cells, we cannot make a definite statement.

The high-confidence list also included 12 transcription factors that employ SCAN domains for co-co assembly (Fig. 4C, right). SCAN domains are leucine-rich, N-terminal motifs composed of five packed  $\alpha$  helices that mediate homo- and hetero-oligomerization of a large family of C2H2 zinc finger proteins by intercalating helix 2 of one monomer between helices 3 and 5 of the opposing monomer.

By comparing the co-co assembly efficiency of these five major dimerization domains, we found that coiled coils conferred the highest (yet very variable) stability to the nascent chain interactions, followed by BTB, BAR, RHD, and SCAN domains (fig. S3E).

Finally, our dataset included two less characterized domains that were significantly enriched (Fig. 4A). The first are STT1 repeats of ubiquitin proteins. This domain mediates homo- and heterodimerization of ubiquitin 1 and 2 (24), both of which were high-confidence candidates that fully exposed the second STT1 repeat (STT1 2) at the assembly onset (fig. S3F). The second, GBD/FH3, are conserved N-terminal regulatory elements in diaphanous-related formins, a protein class involved in nucleation and remodeling of the actin cytoskeleton. The FH3 domain has been implicated in dimerization of the mouse homolog of human *DIAPH1* (25). We found six human formins among our high-confidence proteins; in all cases, the FH3 domain was exposed at assembly onset, suggesting that formins may cotranslationally assemble via the FH3 domain (fig. S3G).

### Co-co assembly is independent of eukaryotic assembly factors

We next examined whether ribosome exposure of co-co assembly-competent nascent chains suffices for disome formation, and whether it could occur outside the eukaryotic folding environment. To investigate these questions, we performed DiSP of *E. coli* that synthesize human lamin C (*LMNA*), one of the mammalian intermediate filaments that were all high-confidence candidates of our DiSP screening. Lamins form homodimers in the cytosol and assemble into higher-order polymers in the nucleus. Dimerization involves the N-terminal rod domain, a long, discontinuous coiled coil that includes three segments (coils 1A, 1B, and 2AB). *LMNA* overexpression generated a disome peak in the RNase-digested lysate (Fig. 5A). DiSP revealed that these disomes were enriched with ribosomes that translate *LMNA* (Fig. 5B), indicating that nascent lamin C can cotranslationally dimerize in bacteria. The minimal length of nascent lamin C mediating the disome

shift in *E. coli* was close to that of the endogenously expressed lamin C in mammalian cells (Fig. 5B). Likewise, overexpression of *DCTN1* generated a disome peak that was enriched with ribosomes exposing the coiled coil of p150<sup>glued</sup>, and the assembly onset was similar to that in human cells (fig. S4A). This observation indicates that co-co assembly of coiled coils is independent of eukaryote-specific assembly factors or mRNA subcellular localization.

To test our hypothesis that the formation of a coiled coil between two nascent chains is minimally required and sufficient to induce disome shifts in bacteria, we used coil 1B of lamin C as a paradigm. First, we employed an established in vivo dimerization assay based on a  $\lambda$  repressor fusion system (26) to show that the isolated 1B efficiently dimerized in *E. coli* (fig. S4B). Second, we performed DiSP to verify that nascent 1B, N-terminally fused to mCherry, efficiently mediated co-co assembly (Fig. 5C, left). Third, we perturbed the periodicity of nonpolar and charged amino acids required for coiled-coil formation of 1B by swapping positions “a” and “e” of the coiled-coil heptameric repeats (1B\*; Fig. 5C, middle). These swaps do not change the overall amino acid composition, the hydrophobicity, or the predicted propensity to form  $\alpha$  helices, but they do eliminate the proficiency of 1B to form a coiled coil (Fig. 5C, insets). In contrast to 1B, the mutated 1B\* did not confer cotranslational disome formation in *E. coli* (Fig. 5C, right), further indicating that DiSP detects productive, in vivo interactions between nascent chains that drive protein oligomer formation.

### Co-co assembly in cis may ensure isoform-specific coiled-coil formation

Lamins A and C are isoforms encoded by the same gene but translated on two alternatively spliced transcripts. Although they share the same N-terminal rod dimerization domain, lamins A and C exclusively form homodimers in vivo (27). How this isoform specificity is achieved in the cellular environment is not known. Co-co assembly may provide a simple answer to this conundrum: Isoform-specific assembly may be achieved by co-co assembly in trans on colocalized mRNAs of the same type [which might segregate in the cytosol, owing to their distinct 3' untranslated regions (UTRs)], or in cis, facilitated by interaction of nascent proteins synthesized by neighboring ribosomes organized in a polysome (Fig. 5D, left).

To distinguish between these possibilities, we generated a heterozygous HEK293-T cell line, in which one *LMNA* allele encodes a C-terminally TwinStrep-tagged lamin C. We performed a series of affinity purification experiments, which revealed that tagged lamin C never copurified the untagged counterpart, even though both proteins are derived from identically spliced mRNAs with identical UTRs

(Fig. 5D, right). This result supports the model that co-co assembly in cis facilitates isoform-specific lamin dimerization in human cells.

### Discussion

In this paper, we provide a comprehensive analysis of cotranslational protein complex assembly mediated by two nascent subunits. The ribosome profiling-based approach that we developed (DiSP) allowed us to identify hundreds of high-confidence candidates and thousands of low-confidence candidates in human cells, revealing co-co assembly as a major route to complex formation.

We decided to include all translocated proteins in the low-confidence list. Many of them are membrane proteins that are often partially or fully resistant to PK but sensitive to puromycin—in particular, small proteins (up to 35 kDa) with multiple annotated TMDs. PK resistance may be conferred by ribosome docking to the translocon that limits the access of PK to the nascent protein. We speculate that docking of ribosomes that closely follow each other in a polysome may spatially organize translocons in the membrane and facilitate homomer assembly.

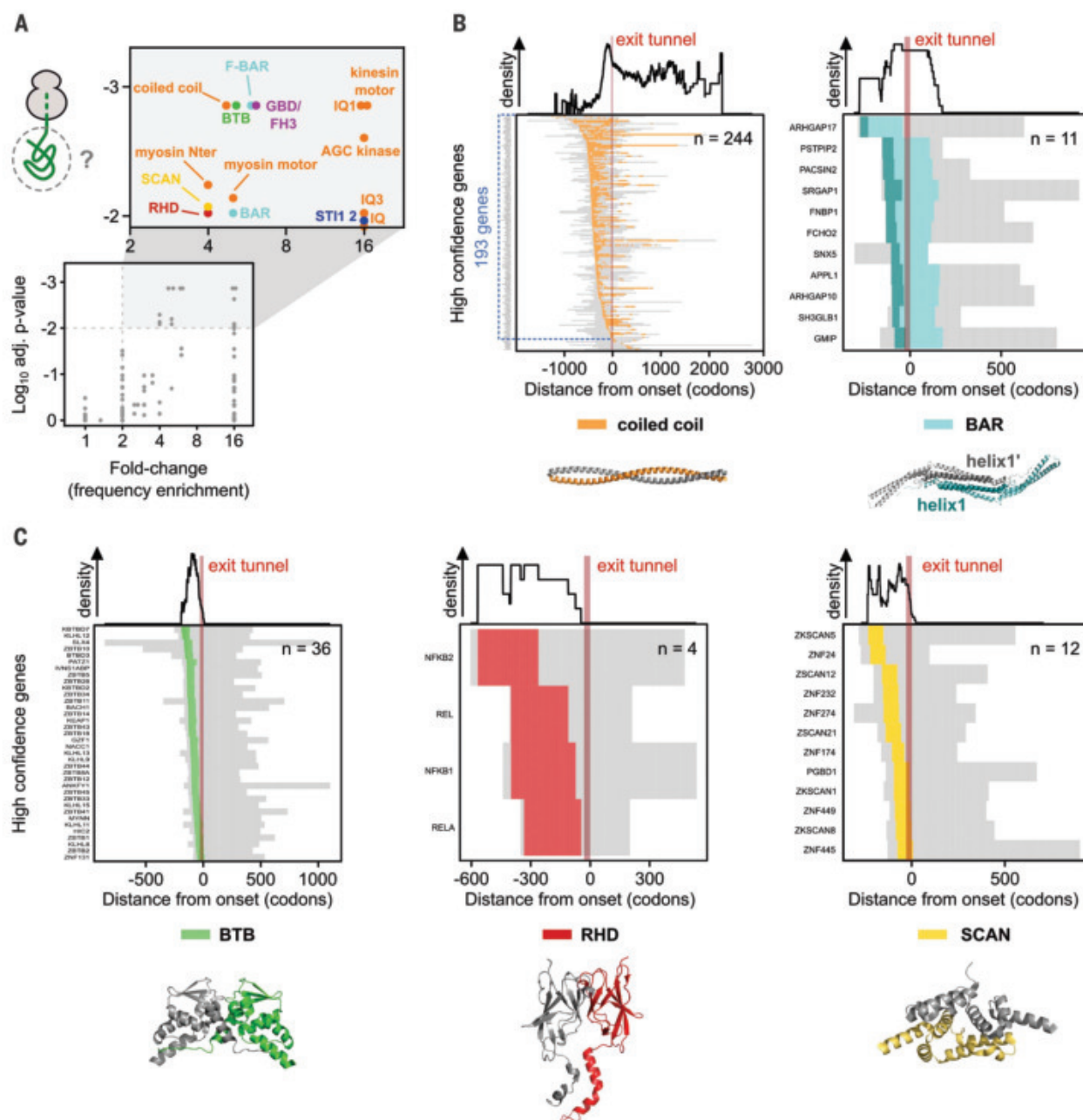
Our data show that predominantly homodimers co-co assemble. We did not find clear evidence that heteromers co-co assemble in trans, because our high-confidence list, in most cases, contained only one subunit of an annotated heteromer. The absence of a known partner subunit may be caused by the less complete structural characterization of heteromeric complexes.

We also did not find clear evidence that the recently described assembly of the TAF6-TAF9 nuclear complex includes nascent chain interactions (14). Both subunits are included in the low-confidence list, but the length of the disome shift and the enrichment efficiency is very different between the two proteins, which is not consistent with a model of co-co assembly in trans.

Co-co assembly of homomers in cis may be facilitated by a generally high ribosome occupancy to ensure close proximity of the interacting nascent chains. In addition, both heteromer assembly (in trans) and homomer assembly (in cis or in trans) may benefit from the slowdown of ribosomes at the onset of assembly, to allow the trailing ribosome translating the same mRNA to catch up or to provide an extended time frame to establish the interaction with another nascent chain translated on a distinct mRNA (13).

We discovered two different types of nascent chain dimerization. The first is a zipper-like formation of coiled coils and BAR domains. For this type, the interaction strength may gradually increase as both nascent chains grow, until enough residues involved in dimerization are ribosome exposed to drive the co-co assembly of stable dimers. The second type of nascent chain dimerization may require the prior folding of a fully emerged,





**Fig. 4. Co-co assembly is coordinated with exposure of five major dimerization domain classes.** (A) Analysis of protein domains on nascent chain segments exposed at assembly onset. The frequency of each domain in the high-confidence class is compared with their general frequency in the proteome (31). We used a Monte Carlo simulation of the null hypothesis to calculate the *P* value (31) and the Benjamini-Yekutieli procedure to correct for multiple testing. The adjusted *P* value is plotted against the respective fold change (frequency enrichment). Domains passing a significance (adjusted *P* ≥ 0.01) and fold change (≥2) threshold are shown in the magnified rectangle and further analyzed. (B) Heatmaps of partially exposed domains: coiled coil (left) and BAR (right). In the heatmaps, nascent

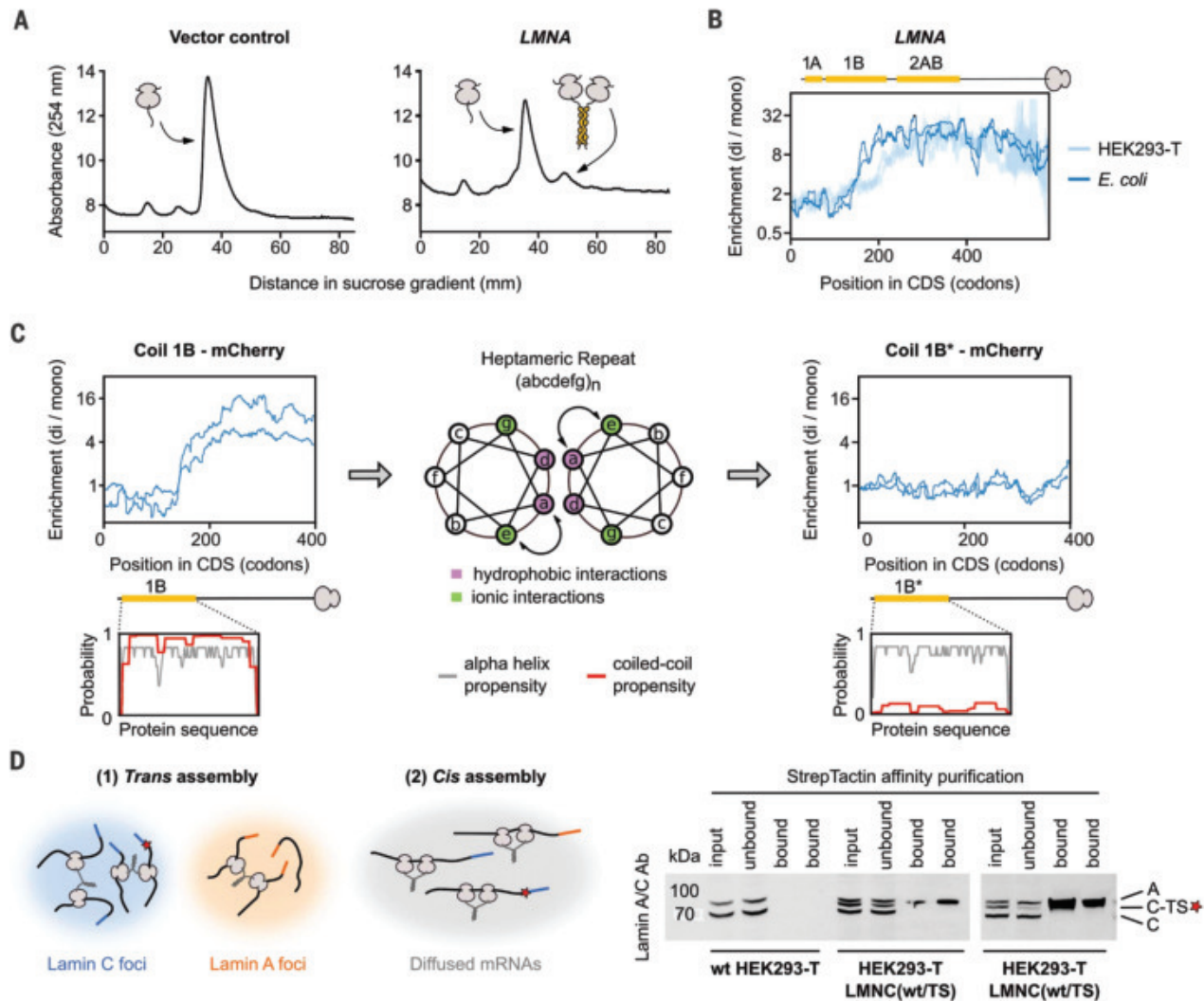
chain segments on the left side of the indicated ribosome exit tunnel (~30 codons, shown by a red bar) are exposed when assembly starts. The subset of genes exposing a coiled-coil segment on the nascent chain at the onset of assembly is highlighted in blue (*n* = 193). Residues forming helix1 of BAR domains are colored dark green in the heatmap and in the exemplary structure. Corresponding domain density profiles are shown atop the heatmaps. Protein Data Bank (PDB) IDs for representative structures: 1D7M (coiled coil) and 3Q0K (BAR). (C) Heatmaps of completely exposed domains: BTB (left), RHD (middle) and SCAN (right). Corresponding domain density profiles are shown atop the heatmaps. PDB IDs for representative structures: 1BUO (BTB), 1K3Z (RHD), and 3LHR (SCAN).

globular interaction domain (a BTB, RHD, or SCAN domain), a feature already reported for co-post assembly (3, 8).

Homodimerization contact regions are evolutionarily selected to be enriched in C-terminal halves of proteins, supposedly to ensure that

folding is not disturbed by the proximity of another identical, incompletely folded subunit (27). Our analysis supports this C-terminal enrichment for most of the human proteome, except for the proteins in our high-confidence list. For the latter proteins, the selective pres-

sure to assemble early apparently outweighs the risk for misfolding of yet-to-be-synthesized C-terminal domains. We speculate that productive folding of the native dimer, beyond co-co assembly, is likely supported by extensive, finely tuned intervention of molecular chaperones.



**Fig. 5. Co-co assembly does not rely on eukaryote-specific factors and facilitates native biogenesis of lamin C homodimers.** (A) Sucrose gradient sedimentation analysis of *E. coli* ribosomes from cells transformed with a control plasmid (left) or a plasmid that encloses human LMNA encoding lamin C (right), lacking the unstructured N-terminal head domain (31). (B) Disome-over-monomer enrichment profile of plasmid-encoded LMNA expressed in *E. coli* (dark blue; two biological replicates), and endogenously expressed LMNA in HEK293-T cells (light blue; two biological replicates). The ribosome-exposed coiled-coil interfaces are indicated by yellow bars. (C) Disome-over-monomer enrichment profiles of LMNA encoding lamin coil 1B (left) or the version of 1B with

positions a and e swapped (1B\*; right) fused N terminally to mCherry and expressed in *E. coli* (two biological replicates). The ribosome-exposed coiled-coil interfaces are indicated by yellow bars. A helical wheel projection shows residue arrangements (a to g) of the heptad repeat (middle). Coiled-coil (red) and  $\alpha$ -helical (gray) probability predictions are shown for both wild-type and mutant 1B (insets). (D) (Left) Hypothetical models of co-co assembly supporting isoform-specific homodimerization. A red star represents the TwinStrep tag (TS). (Right) Affinity purification of tagged lamin C (C-TS) from wild-type (wt) or heterozygous [LMNC(wt/TS)] HEK293-T cells (technical replicates shown). Bands are labeled: A (lamin A), C (lamin C), and C-TS (lamin C-TwinStrep).

Multiple factors may create selective pressure against diffusion-driven assembly and favor co-co assembly: (i) Co-co assembly may increase the efficiency and rate of complex formation. This advantage is most evident for the cis assembly mode in which dimerizing nascent chains are already adjacent within polysomes. (ii) Synthesis-coupled assembly may suppress unproductive interactions and facilitate native folding by limiting the exposure of aggregation-prone dimerization interfaces to the crowded cellular environment. (iii) Cis assembly creates mRNA-specific homomers.

Coiled coils and BTB domains are recurrent dimerization modules in the human proteome, with high potential for non-specific, potentially deleterious heteromeric interactions (28, 29). Such interactions—including those among splicing-derived isoforms that share identical dimerization domains, as in the case of human lamin A and C (27, 28)—would be efficiently prevented in cis assembly. Misassembled subunits that failed to co-co assemble may be recognized by a recently described pathway that specifically detects and eliminates complexes of aberrant composition [dimerization quality control (DQC) (30)]. No-

tably, DQC has been reported as a surveillance mechanism for BTB complexes, but a similar molecular machinery that monitors the composition of other complexes, including coiled coils, may exist. Our proteome-wide study reveals that cotranslational interactions between nascent subunits are a general and efficient strategy to guide the isoform-specific formation of protein complexes.

#### Materials and methods summary

Detailed materials and methods can be found in the supplementary materials.



Human osteosarcoma U2OS (ATCC catalog no. HTB-96), human embryonal kidney HEK293-T (DSMZ catalog no. ACC 635), and *E. coli* Rosetta cells (Novagene) were employed for DiSP experiments.

All ribosome profiling libraries were prepared as described in (20) and sequenced on a NextSeq550 (Illumina) according to the manufacturer's protocol, except for libraries of U2OS samples, which were prepared as described in (8) and sequenced on a HiSeq 2000 (Illumina).

DiSP with PK treatment included incubation of the cell lysates for 30 min at 4°C with the following ratios of PK to total protein: (i) low PK = 1:20,000; (ii) mid PK = 1:6000; (iii) high PK = 1:2000; and (iv) very high PK = 1:200.

DiSP with puromycin omitted cycloheximide from all buffers; cell lysates were incubated for 25 min with 2 mM of puromycin and cross-linked with 0.5% formaldehyde.

## REFERENCES AND NOTES

- G.-W. Li, D. Burkhardt, C. Gross, J. S. Weissman, *Cell* **157**, 624–635 (2014).
- G. Tian *et al.*, *J. Cell Biol.* **138**, 821–832 (1997).
- Y.-W. Shieh *et al.*, *Science* **350**, 678–680 (2015).
- A. Rousseau, A. Bertolotti, *Nat. Rev. Mol. Cell Biol.* **19**, 697–712 (2018).
- L. A. Mingle *et al.*, *J. Cell Sci.* **118**, 2425–2433 (2005).
- M. Pizzinga *et al.*, *J. Cell Biol.* **218**, 1564–1581 (2019).
- B. Hampoelz *et al.*, *Cell* **179**, 671–686.e17 (2019).
- A. Shiber *et al.*, *Nature* **561**, 268–272 (2018).
- G. Kramer, A. Shiber, B. Bukau, *Annu. Rev. Biochem.* **88**, 337–364 (2019).
- C. D. Nicholls, K. G. McLure, M. A. Shields, P. W. K. Lee, *J. Biol. Chem.* **277**, 12937–12945 (2002).
- R. Gilmore, M. C. Coffey, G. Leone, K. McLure, P. W. Lee, *EMBO J.* **15**, 2651–2658 (1996).
- L. Lin, G. N. DeMartino, W. C. Greene, *EMBO J.* **19**, 4712–4722 (2000).
- O. O. Panasenko *et al.*, *Nat. Struct. Mol. Biol.* **26**, 110–120 (2019).
- I. Kamenova *et al.*, *Nat. Commun.* **10**, 1740 (2019).
- N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, J. S. Weissman, *Science* **324**, 218–223 (2009).
- P. Han *et al.*, *Cell Rep.* **31**, 107610 (2020).
- S. D. Redick, J. E. Schwarzbauer, *J. Cell Sci.* **108**, 1761–1769 (1995).
- J. Lu, J. M. Robinson, D. Edwards, C. Deutsch, *Biochemistry* **40**, 10934–10946 (2001).
- F. Liu, D. K. Jones, W. J. de Lange, G. A. Robertson, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 4859–4864 (2016).
- N. J. McGlincy, N. T. Ingolia, *Methods* **126**, 112–129 (2017).
- E. Natan *et al.*, *Nat. Struct. Mol. Biol.* **25**, 279–288 (2018).
- J. Ludwiczak, A. Winski, K. Szczepaniak, V. Alva, S. Dunin-Horkawicz, *Bioinformatics* **35**, 2790–2795 (2019).
- L. Lin, G. N. DeMartino, W. C. Greene, *Cell* **92**, 819–828 (1998).
- D. L. Ford, M. J. Monteiro, *Biochem. J.* **399**, 397–404 (2006).
- R. Rose *et al.*, *Nature* **435**, 513–518 (2005).
- J. C. Hu, E. K. O'Shea, P. S. Kim, R. T. Sauer, *Science* **250**, 1400–1403 (1990).
- T. Kolb, K. Maass, M. Hergt, U. Aebi, H. Herrmann, *Nucleus* **2**, 425–433 (2011).
- Q. Ye, H. J. Worman, *Exp. Cell Res.* **219**, 292–298 (1995).
- G. Schreiber, A. E. Keating, *Curr. Opin. Struct. Biol.* **21**, 50–61 (2011).
- E. L. Mena *et al.*, *Science* **362**, eaap8236 (2018).
- Materials and methods are available as supplementary materials.
- M. D. Young, M. J. Wakefield, G. K. Smyth, A. Oshlack, *Genome Biol.* **11**, R14 (2010).
- ilia-kats, ilia-kats/RiboSeqTools: v0.1, Zenodo (2020); doi: 10.5281/ZENODO.4016066.

## ACKNOWLEDGMENTS

We thank all members of B.B.'s laboratory for discussions and advice; D. Coombs for help with optimization of ribosome separation on sucrose gradients; U. Friedrich for help with establishing pipelines for processing ribosome profiling data;

S. Anders for valuable advice concerning development of DiSP data analysis tools; the ZMBH Flow Cytometry & FACS Core Facility, the DKFZ Sequencing Core Facility, and the DKFZ Vector and Clone Repository for support of experimental work. M.B., K.F., and J.S. are members of the Heidelberg Biosciences International Graduate School (HBIGS). **Funding:** M.B. and K.F. were supported by a HBIGS Ph.D. fellowship. M.B. was additionally supported by a Boehringer Ingelheim Fonds (BIF) Ph.D. fellowship. F.W. received funding from the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie grant 745798. This work was supported by the Helmholtz-Gemeinschaft [DKFZ NCT3.0 Integrative Project in Cancer Research (DysregPT\_Bukau 1030000008 G783)], the Deutsche Forschungsgemeinschaft (SFB 1036), the European Research Council [ERC Advanced Grant (743118)], and the Klaus Tschira Foundation. Work in the Tans laboratory was supported by the Netherlands Organization for Scientific Research (NWO). **Authors contributions:** Conceptualization: M.B., K.F., F.W., J.S., S.T., B.B., and G.K. Methodology: M.B., K.F., J.J.A., B.B., and G.K. Investigation: M.B. and K.F. Software: I.K., F.T., M.B., and K.F. Formal analysis, Data curation, and Visualization: M.B., K.F., I.K., F.T., B.B., and G.K. Writing – original draft: M.B., K.F., I.K., and G.K. Writing – review &

editing: all authors. Supervision: S.T., B.B., and G.K. **Competing interests:** All authors declare no competing interests. **Data and materials availability:** All sequencing data reported in this study are available at GEO under accession number GSE151959. Explicit Julia code is available as supplementary material; explicit R code will be made available upon request. Data analysis of ribosome profiling datasets was performed with RiboSeqTools [available at <https://github.com/ilia-kats/RiboSeqTools> and Zenodo (33)].

## SUPPLEMENTARY MATERIALS

[science.sciencemag.org/content/371/6524/57/suppl/DC1](https://science.sciencemag.org/content/371/6524/57/suppl/DC1)  
Materials and Methods  
Figs. S1 to S4  
Tables S1 to S5  
References (34–56)  
MDAR Reproducibility Checklist  
Custom Julia Scripts 1 to 3

16 June 2020; accepted 27 October 2020  
10.1126/science.abc7151

## REPORTS

### CELL CYCLE

# A tripartite mechanism catalyzes Mad2-Cdc20 assembly at unattached kinetochores

Pablo Lara-Gonzalez<sup>1,2,3,\*</sup>, Taekyung Kim<sup>1,2,3,†</sup>, Karen Oegema<sup>1,2,3</sup>, Kevin Corbett<sup>2,3</sup>, Arshad Desai<sup>1,2,3,\*</sup>

During cell division, kinetochores couple chromosomes to spindle microtubules. To protect against chromosome gain or loss, kinetochores lacking microtubule attachment locally catalyze association of the checkpoint proteins Cdc20 and Mad2, which is the key event in the formation of a diffusible checkpoint complex that prevents mitotic exit. We elucidated the mechanism of kinetochore-catalyzed Mad2-Cdc20 assembly with a probe that specifically monitors this assembly reaction at kinetochores in living cells. We found that catalysis occurs through a tripartite mechanism that includes localized delivery of Mad2 and Cdc20 substrates and two phosphorylation-dependent interactions that geometrically constrain their positions and prime Cdc20 for interaction with Mad2. These results reveal how unattached kinetochores create a signal that ensures genome integrity during cell division.

**D**uring cell division, the centromere regions of replicated chromosomes assemble kinetochores, mechanical interfaces that couple sister chromatids to spindle microtubules (1). Kinetochores also serve as signaling hubs that monitor their own attachment status; when unattached, they delay anaphase onset by producing an inhibitor of the anaphase-promoting complex/cyclosome (APC/C), the E3 ubiquitin ligase that promotes exit from mitosis (2). Prior work has shown that the essential role of kinetochores in APC/C inhibitor production is to catalyze formation

of a complex between the checkpoint proteins Mad2 and Cdc20 (3). Once formed, the Mad2-Cdc20 complex rapidly binds Mad3 (BubR1) and Bub3 in the cytosol to form the mitotic checkpoint complex that inhibits the APC/C. Central to complex formation is the ability of Mad2 to adopt two conformational states: an open, free form and a closed ligand-bound form (4, 5). The Mad2 ligands in checkpoint signaling are Mad1, a dimeric coiled-coil protein with a folded C-terminal domain, and Cdc20 (4, 5). The Mad1-Mad2 complex is present throughout the cell cycle, whereas Mad2 assembly onto Cdc20 is kinetically disfavored and requires catalysis by unattached kinetochores (Fig. 1A) (6, 7). In the current model of checkpoint signaling, a stable kinetochore-anchored complex of Mad1 with closed Mad2 recruits an open conformer of Mad2 from the cytoplasm via asymmetric dimerization; this open Mad2 is subsequently linked to Cdc20 (Fig. 1A) (4, 5). How kinetochores overcome the kinetic barrier to Mad2-Cdc20 association

<sup>1</sup>Section of Cell and Developmental Biology, Division of Biological Sciences, University of California, San Diego, La Jolla, CA, USA. <sup>2</sup>Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA, USA. <sup>3</sup>Ludwig Institute for Cancer Research, San Diego Branch, 9500 Gilman Drive, La Jolla, CA, USA.

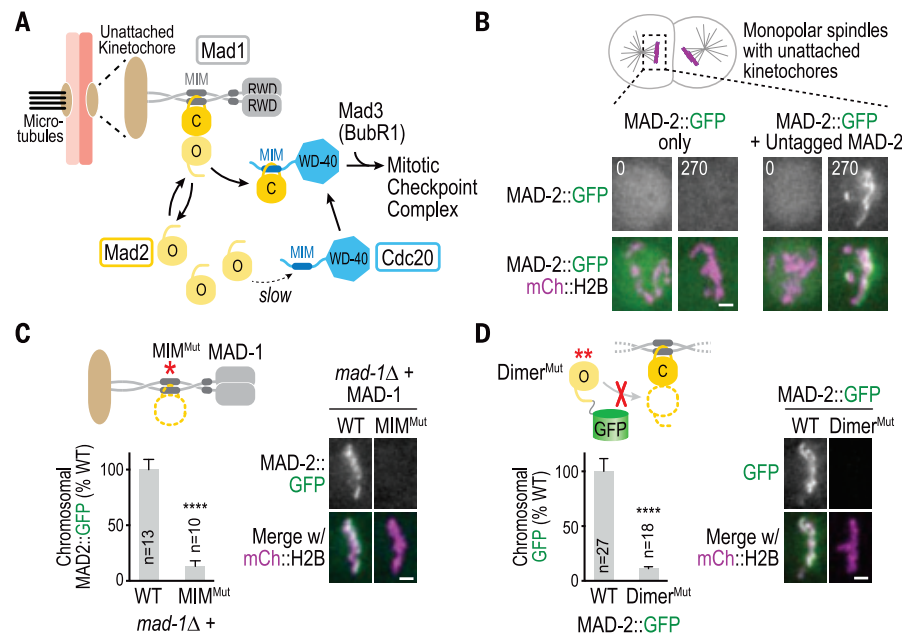
\*Corresponding author. Email: [abdesai@ucsd.edu](mailto:abdesai@ucsd.edu) (A.D.); [plgonzalez@ucsd.edu](mailto:plgonzalez@ucsd.edu) (P.L.-G.)

†Present address: Department of Biology Education, Pusan National University, Busan 26241, Republic of Korea.

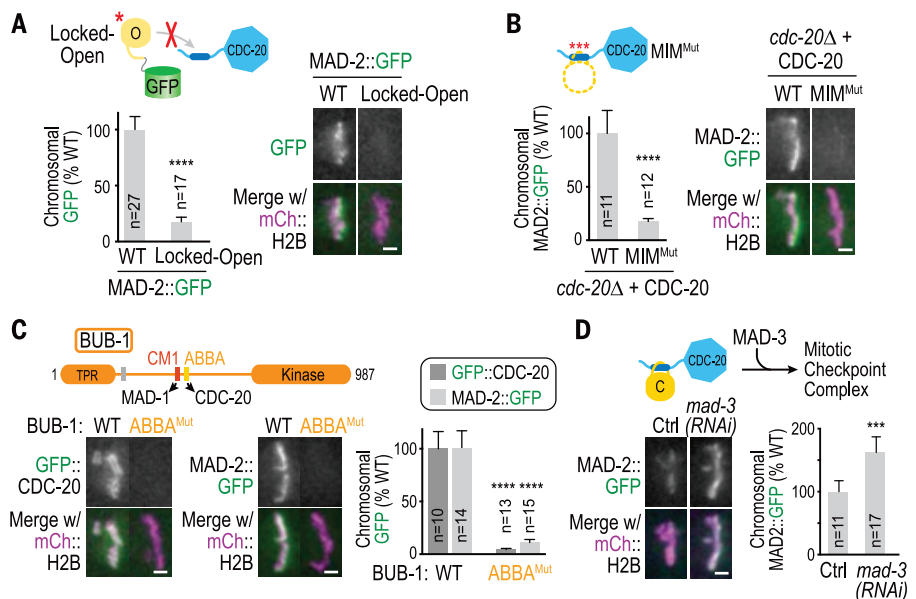
is not understood. Reconstitutions have highlighted a role for Mad1 phosphorylation (6, 8), but it is unclear how this phosphorylation acts, whether it is sufficient for catalysis in vivo, and why catalysis occurs specifically at kinetochores.

The presence of two kinetochore-localized pools of Mad2 (9) has prevented selective monitoring of the cycling pool that becomes linked to Cdc20, limiting efforts to unravel the steps involved in Cdc20-Mad2 assembly at kinetochores. In an effort to endogenously tag MAD-2 in *Caenorhabditis elegans* with green fluorescent protein (GFP), we fortuitously generated a C-terminal MAD-2::GFP fusion that specifically visualizes the cycling pool of MAD-2 at kinetochores in living embryos. MAD-2::GFP localization to unattached kinetochores was compromised when it was the sole source of MAD-2 (Fig. 1B and fig. S1A). Embryos expressing only MAD-2::GFP were defective in checkpoint signaling (10), which initially suggested loss of MAD-2 function. However, MAD-2::GFP localized to kinetochores when expressed together with untagged MAD-2 (Fig. 1B and fig. S1B), which implies that MAD-2::GFP can be recruited to kinetochores through dimerization with untagged MAD-2 bound to MAD-1. Consistent with this model, engineering MAD-1 to selectively disrupt MAD-2 binding eliminated MAD-2::GFP kinetochore localization, as did disrupting the dimerization interface on MAD-2::GFP (Fig. 1, C and D, fig. S1C, and fig. S2). Thus, MAD-2::GFP, when expressed alongside untagged MAD-2, selectively monitors the MAD-2 that cycles through kinetochores during checkpoint signaling (fig. S3).

The cycling pool of Mad2 localizes to kinetochores through asymmetric dimerization with Mad2 stably bound to Mad1 and then undergoes conformational conversion to a closed Cdc20-bound state (3, 11). To test which of these events is required to observe kinetochore-localized MAD-2::GFP, we generated a locked-open mutant that exhibits normal dimerization but cannot undergo open-to-closed conversion (Fig. 2A and fig. S2, A and B). Locked-open MAD-2::GFP was not detected at kinetochores, indicating that stable binding of cycling MAD-2 at kinetochores depends not only on dimerization with untagged MAD-2 but also on the conversion of MAD-2 to its CDC-20-bound closed form. In support of this model, both CDC-20's MAD-2 interaction motif (10) and the ABBA motif of BUB-1 that recruits CDC-20 to kinetochores (12, 13) were required for MAD-2::GFP kinetochore localization (Fig. 2, B and C, and fig. S4A). These results indicate that the MAD-2::GFP probe reveals MAD-2-CDC-20 complex formation at kinetochores. Consistent with this idea, locked-open MAD-2::GFP dominantly inhibited checkpoint signaling in the presence of untagged MAD-2 (fig. S2C).

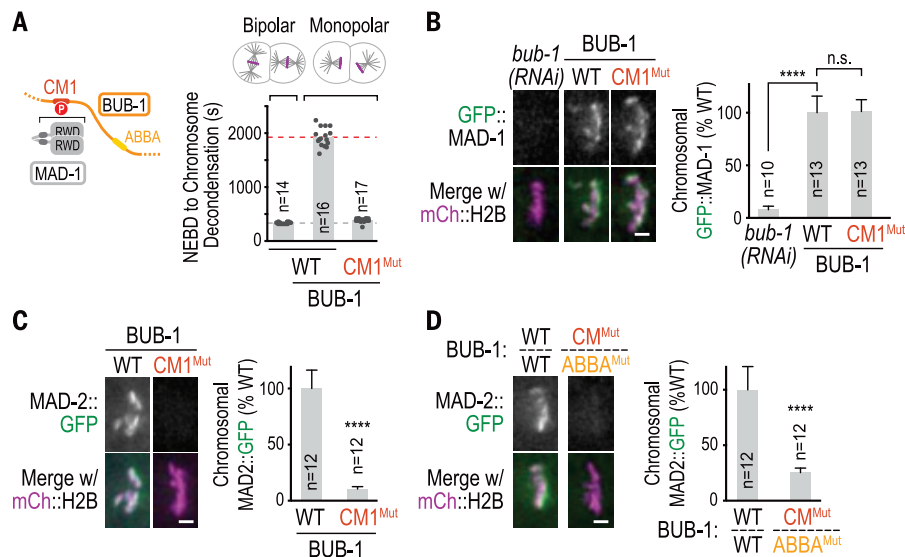


**Fig. 1. A probe that selectively visualizes the cycling pool of Mad2 at kinetochores.** (A) Schematic of spindle checkpoint signaling. O and C indicate open and closed Mad2 conformers, respectively. See text for details. (B) Localization of MAD-2::GFP on its own (left) or in the presence of untagged MAD-2 (right) on monopolar spindles in two-cell *C. elegans* embryos (see fig. S1A for experimental details). All subsequent localization analysis of MAD-2::GFP was on monopolar spindles in the presence of untagged MAD-2. (C and D) MAD-2::GFP localization after introduction of structure-guided mutations in MAD-1's MAD-2 interaction motif (MIM) (C) and in MAD-2's dimerization interface (D). WT, wild type. Red asterisks represent engineered mutations (table S2). In this and all subsequent figures,  $n$  = number of embryos imaged and quantified; error bars denote 95% confidence interval. \*\*\*\* $P$   $\leq$  0.0001 (Mann-Whitney test). Scale bars, 2  $\mu$ m.

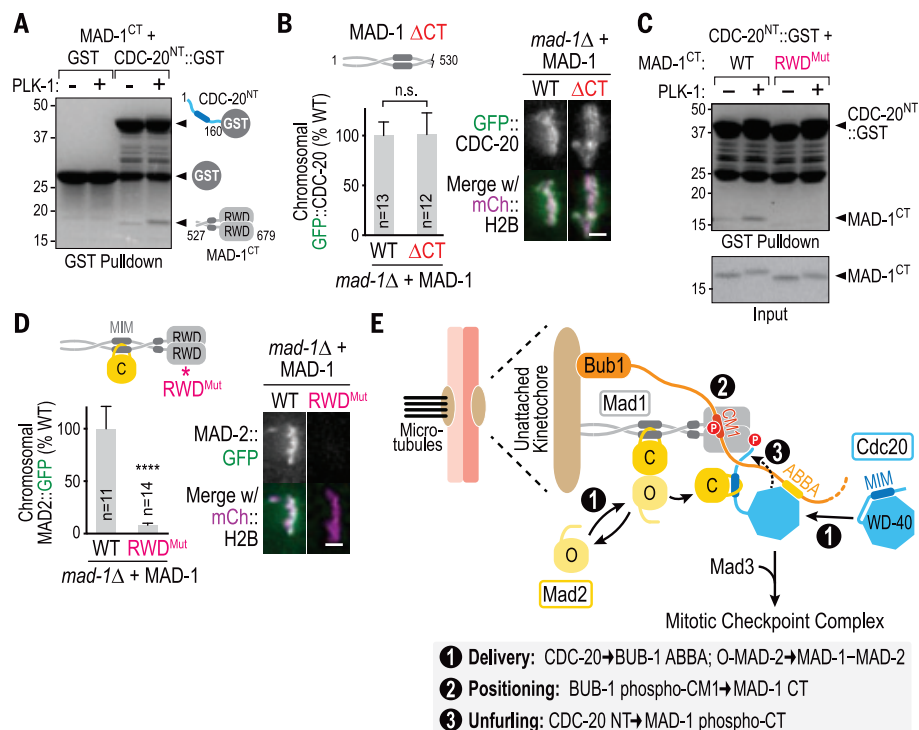


**Fig. 2. MAD-2::GFP reveals complex formation with CDC-20 at the kinetochore in living cells.** (A) Left: Schematic of a locked-open version of MAD-2, which exhibits normal dimerization but is unable to bind ligands, including CDC-20 (fig. S2A). Right: Comparison of wild-type and locked-open MAD-2::GFP localization; data for wild type are the same as in Fig. 1D. (B) MAD-2::GFP localization in the presence of wild-type or MAD-2 interaction motif-mutant (MIM<sup>Mut</sup>) CDC-20. (C) Top: Schematic of BUB-1, whose ABBA motif binds to the CDC-20 WD40 domain. Bottom: GFP::CDC-20 and MAD-2::GFP localization in the presence of wild-type or ABBA-mutant BUB-1. (D) MAD-2::GFP localization after RNA interference (RNAi)-mediated depletion of MAD-3. Red asterisks in (A) and (B) represent engineered mutations (table S2). \*\*\*\* $P$   $\leq$  0.001, \*\*\*\* $P$   $\leq$  0.0001 (Mann-Whitney test). Scale bars, 2  $\mu$ m.





**Fig. 3. The conserved motif of Bub1 positions Cdc20 bound to Bub1 to engage with Mad2.** (A) Left: Schematic of Bub1 conserved motif (CM1), when phosphorylated, mediating an interaction with the C terminus of Mad1. Right: Checkpoint signaling analysis, conducted by comparing mitotic duration in two-cell embryos with bipolar (gray dashed line) versus monopolar (red dashed line) spindles (13). (B and C) Localization of GFP::MAD-1 (B) and MAD-2::GFP (C) in the indicated conditions. (D) MAD-2::GFP localization in a heterozygous *bub-1* mutant, with one allele expressing the CM1 mutant and the other allele expressing the ABBA mutant; *bub-1* mutations were engineered at the endogenous locus. \*\*\*\* $P \leq 0.0001$  (Mann-Whitney test); n.s., not significant. Scale bars, 2  $\mu$ m.



**Fig. 4. Phosphorylation-regulated interaction of Mad1 and Cdc20 promotes Cdc20 association with Mad2.** (A) Coomassie-stained protein gel showing PLK-1 phosphorylation-promoted interaction of the MAD-1 C terminus with the CDC-20 N terminus. (B) CDC-20 localization in wild-type versus  $\Delta$ CT MAD-1. (C and D) CDC-20 interaction (C), analyzed as in (A), and MAD-2::GFP localization (D) for wild-type versus RWD-mutant MAD-1. Purple asterisk represents engineered mutations (table S2). See fig. S7D for details on residues mutated. \*\*\*\* $P \leq 0.0001$  (Mann-Whitney test). Scale bars, 2  $\mu$ m. (E) Model of the tripartite mechanism that catalyzes localized Mad2-Cdc20 assembly at the kinetochore.

If MAD-2::GFP monitors the rate-limiting step of mitotic checkpoint complex assembly—formation of the MAD-2-CDC-20 complex at kinetochores—then MAD-2::GFP localization should be unaffected by removal of MAD-3 (fig. S4, B and C). In fact, MAD-3 depletion increased MAD-2::GFP signal at kinetochores (Fig. 2D), indicating that capture by MAD-3 may aid release of MAD-2::GFP-CDC-20 complexes from kinetochores.

Adjacent to its Cdc20-binding ABBA motif, Bub1 contains a conserved motif (CM1) that, when phosphorylated, mediates an interaction with the Mad1 C terminus (8, 14, 15) (fig. S5A). Although Mad1-Mad2 kinetochore localization depends on phosphorylated BUB-1 CM1 in some species (14–16), its importance in checkpoint signaling is independent of Mad1-Mad2 localization (16), suggesting a second, more conserved function. In *C. elegans*, mutating BUB-1 CM1 had no effect on MAD-1 kinetochore localization and only a mild effect on CDC-20 localization but eliminated the MAD-2::GFP signal at kinetochores and fully inhibited checkpoint signaling (Fig. 3, A to C, and fig. S5, B to D). Consistent with phosphorylation being key to CM1 function in MAD-2-CDC-20 assembly, mutation of a single conserved threonine (Thr<sup>407</sup>) in CM1 eliminated MAD-2::GFP signal at kinetochores and inhibited checkpoint signaling (fig. S5, E and F). Phosphorylated CM1 interacts with the Mad1 C terminus (8, 15), and deletion of the MAD-1 C terminus also eliminated MAD-2-CDC-20 assembly at kinetochores without perturbing MAD-1 localization or MAD-1-MAD-2 complex formation (fig. S6). Thus, delivering MAD-2 and CDC-20 to kinetochores, via their respective docking sites on the MAD-1-MAD-2 complex and BUB-1, appears insufficient to catalyze complex formation; instead, phosphorylation of the BUB-1 CM1 motif and its subsequent interaction with the MAD-1 C terminus are also required. Consistent with the idea that physical proximity of BUB-1's MAD-1 and CDC-20 binding motifs is critical to mediate MAD-2-CDC-20 assembly, *bub-1* alleles with separate mutations in CM1 and ABBA failed to complement one another (Fig. 3D).

The above data highlight a critical role for the BUB-1 ABBA and CM1 motifs in recruiting and positioning CDC-20 for assembly with MAD-2. However, they do not address why MAD-2-CDC-20 association is disfavored in the cytoplasm, nor how kinetochores overcome this barrier. Phosphorylation of the Mad1 C terminus facilitates interaction with the N-terminal region of Cdc20 and promotes Mad2-Cdc20 complex formation in vitro (8). In humans, Mad1 is phosphorylated by Mps1, but *C. elegans* lacks Mps1, and its checkpoint function is provided by Polo-like kinase 1 (PLK-1) (17). Phosphorylation by PLK-1 of the MAD-1 C terminus promoted its interaction

with the CDC-20 N terminus in vitro (Fig. 4A and fig. S7, A and B). This interaction was not required to recruit CDC-20 to kinetochores, because CDC-20 localized normally in the absence of the MAD-1 C terminus (Fig. 4B). In solution, the motif in the Cdc20 N terminus that interacts with Mad2 is masked, potentially by an interaction with its C-terminal WD40 domain (fig. S7C) (18, 19). By interacting with its N terminus, phosphorylated Mad1 may expose Cdc20's Mad2 interaction motif and, in conjunction with the local concentration and positioning mechanisms described above, drive its assembly onto Mad2 at kinetochores. This idea was tested using MAD-1 mutants that prevent its phosphorylation-stimulated interaction with the CDC-20 N terminus; three residues on the MAD-1 RWD domain were targeted for this purpose (Fig. 4C and fig. S7, A and D) (20). The MAD-1 RWD mutant did not exhibit phosphorylation-dependent binding to the CDC-20 N terminus in vitro; when introduced in vivo, it eliminated formation of the MAD-2::CDC-20 complex at kinetochores and was deficient in checkpoint signaling (Fig. 4D and fig. S7, E to G). One of the critical phosphorylation sites that mediates interaction with Cdc20 in human Mad1 (Thr<sup>716</sup>) (6, 8) is not conserved outside of vertebrates (fig. S8A). Analysis of six putative phosphosites in the *C. elegans* MAD-1 C terminus indicated that one conserved residue (Thr<sup>653</sup>) was important for MAD-2-CDC-20 association at kinetochores and for checkpoint signaling (fig. S7, F and G, and fig. S8). These data support a model in which phosphorylation of Mad1 at kinetochores promotes an interaction with Cdc20 that exposes its Mad2 interaction motif (6, 8, 21).

Visualization of kinetochore-localized Mad2-Cdc20 assembly, together with component engineering, reveals how the spindle checkpoint signal is generated at unattached kinetochores (Fig. 4E). Localized signal generation involves three steps: (i) delivery of Cdc20 (by the Bub1 ABBA motif) and Mad2 (through dimerization with the Mad1-Mad2 template); (ii) positioning of Cdc20 by interaction of the Bub1 CM1 motif with the Mad1 C terminus; and (iii) unfurling of the Cdc20 N terminus by interaction with phosphorylated Mad1, exposing its Mad2 interaction motif. The requisite multipartite interaction network explains the specific assembly of Mad2-Cdc20 signaling complexes at unattached kinetochores. Within this network, Bub1 plays a key "matchmaker" role, mediating Cdc20 recruitment (via the ABBA motif) and positioning (via the CM1 motif interacting with Mad1) at unattached kinetochores. Catalysis additionally requires phosphorylation of Mad1, which likely serves to unfurl Cdc20 and expose its Mad2 interaction motif. Thus, the simultaneous interaction of Cdc20 with Bub1 and Mad1 is critical

for overcoming the kinetic barrier to its association with Mad2. Our conclusions are in broad agreement with those from a biochemical reconstitution of the spindle checkpoint using purified components (21). These two studies address the long-standing question of how kinetochores direct localized production of signaling complexes that act as critical guardians of the genome during cell division.

## REFERENCES AND NOTES

1. A. Musacchio, A. Desai, *Biology* **6**, 5 (2017).
2. D. Barford, *Curr. Opin. Struct. Biol.* **61**, 86–97 (2020).
3. K. D. Corbett, *Prog. Mol. Subcell. Biol.* **56**, 429–455 (2017).
4. X. Luo, H. Yu, *Structure* **16**, 1616–1625 (2008).
5. A. Musacchio, *Curr. Biol.* **25**, R1002–R1018 (2015).
6. A. C. Faesen et al., *Nature* **542**, 498–502 (2017).
7. G. Fang, *Mol. Biol. Cell* **13**, 755–766 (2002).
8. Z. Ji, H. Gao, L. Jia, B. Li, H. Yu, *eLife* **6**, e22513 (2017).
9. J. V. Shah et al., *Curr. Biol.* **14**, 942–952 (2004).
10. P. Lara-Gonzalez et al., *Dev. Cell* **51**, 313–325.e10 (2019).
11. G. Zhang, J. Nilsson, *Cell Cycle* **17**, 1087–1091 (2018).
12. B. Di Fiore et al., *Dev. Cell* **32**, 358–372 (2015).
13. T. Kim et al., *Genes Dev.* **31**, 1089–1094 (2017).
14. C. Klebig, D. Korinith, P. Meraldi, *J. Cell Biol.* **185**, 841–858 (2009).
15. N. London, S. Biggins, *Genes Dev.* **28**, 140–152 (2014).
16. S. Heinrich et al., *EMBO Rep.* **15**, 291–298 (2014).
17. J. Espeut et al., *Cell Rep.* **12**, 58–65 (2015).

18. Y. Zhang, E. Lees, *Mol. Cell. Biol.* **21**, 5190–5199 (2001).
19. J. S. Han et al., *Mol. Cell* **51**, 92–104 (2013).
20. S. Kim, H. Sun, D. R. Tomchick, H. Yu, X. Luo, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 6549–6554 (2012).
21. V. Piano et al., *Science* **371**, 67–71 (2021).

## ACKNOWLEDGMENTS

We thank K.-Y. Lee for protein purification advice; J. Woodruff for PLK-1; A. Musacchio and V. Piano for discussing unpublished results; the Caenorhabditis Genetic Collection (CGC) for strains; and J. Houston, R. Green, and J. S. Gomez-Cavazos for comments on the manuscript. **Funding:** NIH R01GM074215 (A.D.); NIH R01GM104141 (K.C.); Ludwig Institute for Cancer Research (A.D.); S10 OD021724 (UCSD mass spectrometry); National Research Foundation of Korea 2020R1C1C1008696 (T.K.). **Author contributions:** P.L.-G. and A.D. initiated the project; P.L.-G. performed the majority of experiments; T.K. generated and initially characterized the CM1 mutant; K.C. guided structure-based and biochemical experiments; and P.L.-G., K.O., K.C., and A.D. prepared 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

science.sciencemag.org/content/371/6524/64/suppl/DC1  
Materials and Methods  
Figs. S1 to S8  
Tables S1 to S5  
References (22–32)

8 April 2020; accepted 17 November 2020  
10.1126/science.abc1424

## CELL CYCLE

# CDC20 assists its catalytic incorporation in the mitotic checkpoint complex

Valentina Piano<sup>1\*</sup>, Amal Alex<sup>1</sup>, Patricia Stege<sup>1</sup>, Stefano Maffini<sup>1</sup>, Gerardo A. Stoppiello<sup>1</sup>, Pim J. Huis in 't Veld<sup>1</sup>, Ingrid R. Vetter<sup>1</sup>, Andrea Musacchio<sup>1,2\*</sup>

Open (O) and closed (C) topologies of HORMA-domain proteins are respectively associated with inactive and active states of fundamental cellular pathways. The HORMA protein O-MAD2 converts to C-MAD2 upon binding CDC20. This is rate limiting for assembly of the mitotic checkpoint complex (MCC), the effector of a checkpoint required for mitotic fidelity. A catalyst assembled at kinetochores accelerates MAD2::CDC20 association through a poorly understood mechanism. Using a reconstituted SAC system, we discovered that CDC20 is an impervious substrate for which access to MAD2 requires simultaneous docking on several sites of the catalytic complex. Our analysis indicates that the checkpoint catalyst is substrate assisted and promotes MCC assembly through spatially and temporally coordinated conformational changes in both MAD2 and CDC20. This may define a paradigm for other HORMA-controlled systems.

**F**aithful chromosome segregation is prerequisite to genome integrity during mitosis. The spindle-assembly checkpoint (SAC) delays anaphase onset until all sister chromatids are attached through their kinetochores to microtubule fibers connected to opposite poles of the mitotic spindle (biorientation) (1). The kinetochore is the sensor of biorientation. A single unattached or incor-

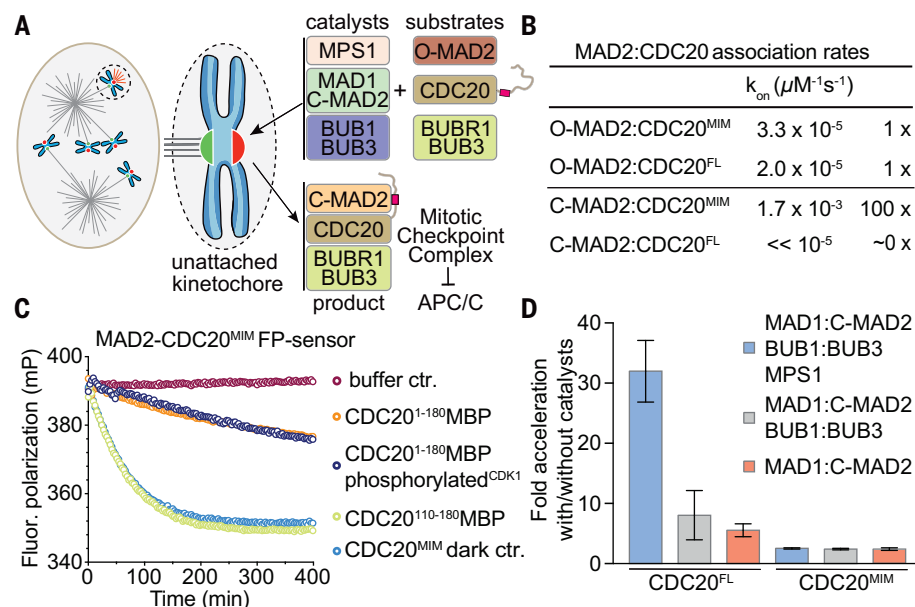
rectly attached kinetochore suffices to trigger the SAC (2). The SAC effector, the MCC, inhibits the anaphase-promoting complex (also called cyclosome or APC/C), the E3 ubiquitin ligase required for sister chromatid separation and mitotic exit (Fig. 1A) (3, 4). The MCC comprises CDC20, C-MAD2, and BUBR1:BUB3 (5, 6). The first step in MCC assembly is a spontaneous but very slow reaction at cell-like concentrations (7–10), where the C-terminal region of MAD2 (1), known as the safety belt (8, 11, 12), transitions from the O- to C-conformation to latch onto the MAD2 interaction motif (MIM) of CDC20 (fig. S1A and B) (13). MAD1:C-MAD2, BUB1:BUB3, and MPS1 (collectively indicated

<sup>1</sup>Department of Mechanistic Cell Biology, Max Planck Institute of Molecular Physiology, 44227 Dortmund, Germany. <sup>2</sup>Centre for Medical Biotechnology, Faculty of Biology, University Duisburg-Essen, 45141 Essen, Germany. \*Corresponding author. Email: valentina.piano@mpi-dortmund.mpg.de (V.P.); andrea.musacchio@mpi-dortmund.mpg.de (A.M.)



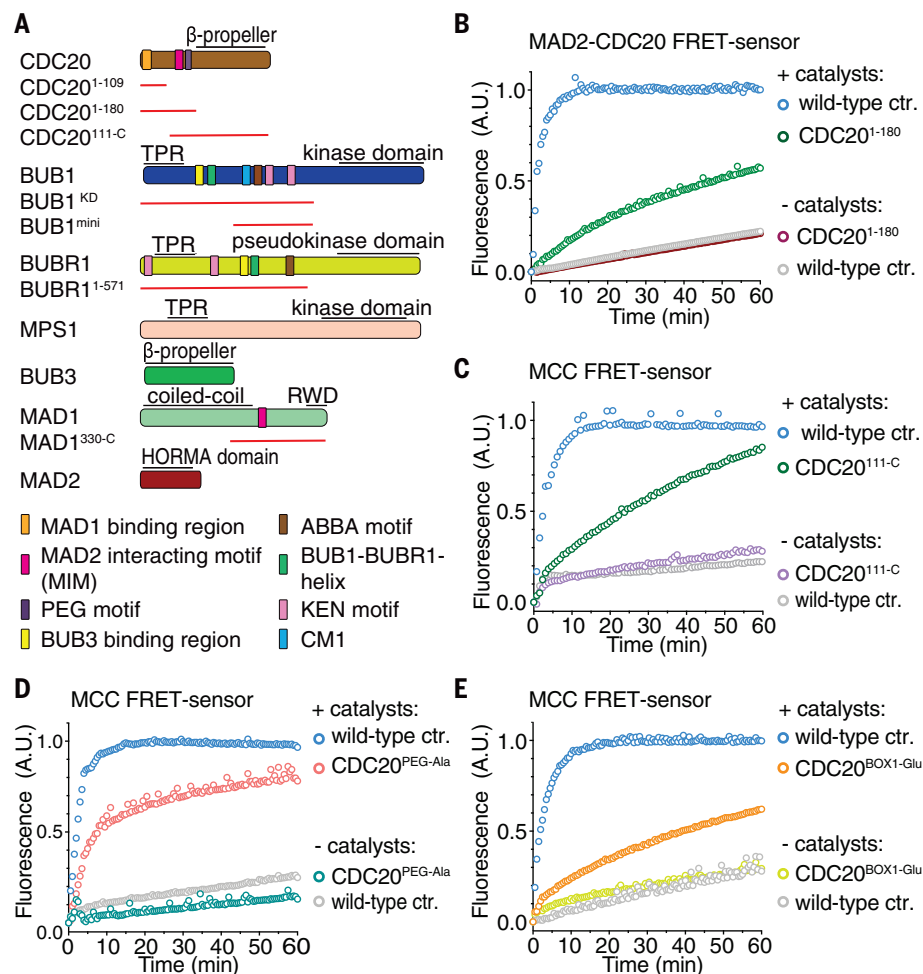
**Fig. 1. Catalysis of O- to C-MAD2 conversion requires CDC20.** (A) Unattached kinetochores recruit catalysts and substrates required to assemble MCC and delay anaphase onset.

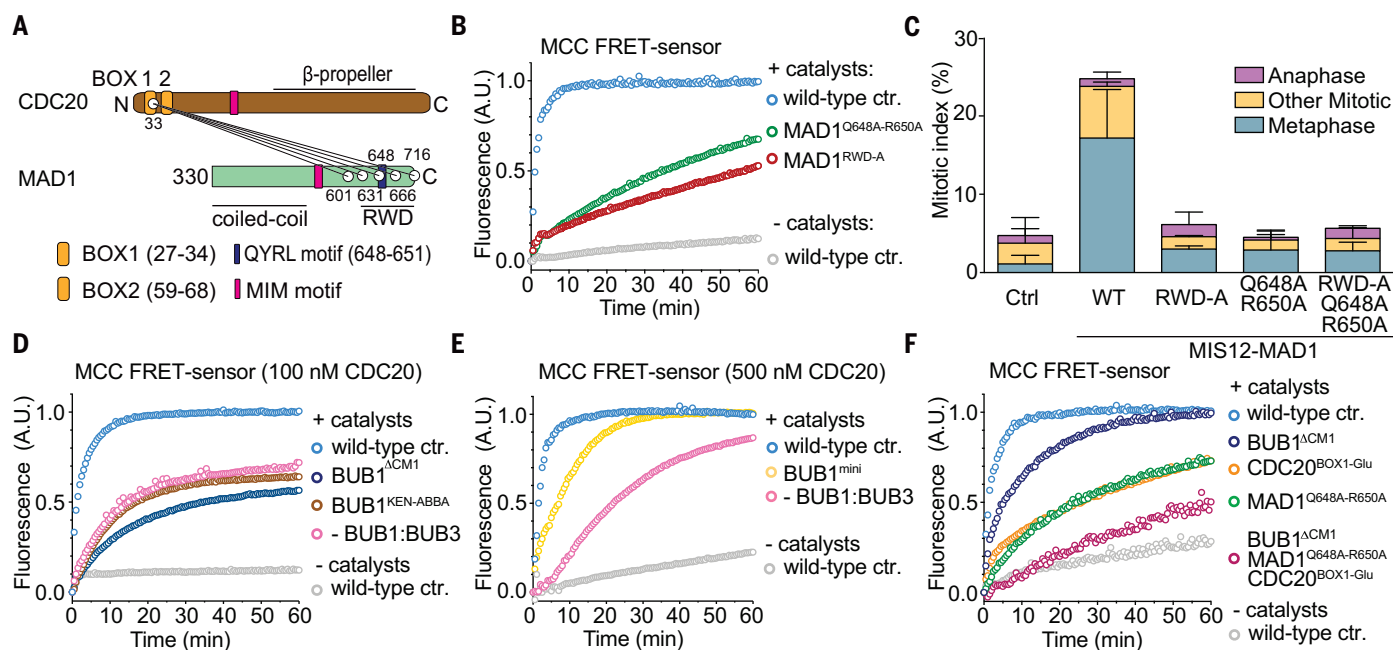
(B) Comparison of second-order association constants ( $k_{on}$ ) for the interaction of CDC20<sup>MIM</sup> or CDC20<sup>FL</sup> with O- or C-MAD2. (C) Displacement assay with preassembled FAM-labeled C-MAD2: CDC20<sup>MIM</sup> measuring the decrease of fluorescence polarization when competing with the indicated CDC20-unlabeled constructs. Throughout, time-dependent changes in fluorescence [fluorescence polarization (FP) or fluorescence resonance energy transfer (FRET)] signal reflect single measurements representative of at least three independent technical replicates. (D) Comparison of the maximal acceleration provided by the catalysts on the association between O-MAD2 and CDC20<sup>MIM</sup> or CDC20<sup>FL</sup>. Bars indicate  $\pm$  SEM of three independent technical replicates.



**Fig. 2. CDC20 catalyzes its incorporation in the MCC.** (A) Schematic representation of the constructs used in this study (see also tables S1 and S2).

(B) FRET assay monitoring the association of MAD2:CDC20<sup>FL</sup> (light blue) and MAD2:CDC20<sup>1-180</sup> (dark green). Unless otherwise specified, here and in other figures depicting FRET assays, the y axis represents the normalized emission intensity of the FRET acceptor indicated as "Fluorescence (A.U.)," where A.U. is arbitrary units. (C) FRET assay monitoring the catalytic assembly of the MCC in the presence of CDC20<sup>FL</sup> (light blue) or CDC20<sup>111-C</sup> (dark green). (D) FRET assay monitoring MCC formation in the presence of CDC20<sup>FL</sup> (light blue) or CDC20<sup>BOX1-Glu</sup> (salmon). (E) FRET assay monitoring MCC formation in the presence of CDC20<sup>FL</sup> (light blue) or CDC20<sup>BOX1-Glu</sup> (orange).





**Fig. 3. BUB1:BUB3 is a scaffold favoring the MAD1:CDC20 interaction.**

(A) XL-MS of the interaction of the CDC20 N-terminal region with MAD1<sup>CTD</sup>. (B) FRET assay monitoring MCC catalytic assembly with wild-type MAD1 (light blue), MAD1<sup>Q648A-R650A</sup> (dark green), or MAD1<sup>RWD-A</sup> (red) (9). (C) HeLa cells transfected with mCherry, mCherry-MIS12-MAD1 wild type (WT), mCherry-MIS12-MAD1<sup>RWD-A</sup> (RWD-A), mCherry-MIS12-MAD1<sup>Q648A-R650A</sup> (Q648A-R650A), or mCherry-MIS12-MAD1<sup>RWD-A-Q648A-R650A</sup> (RWD-A-Q648A-R650A). Graphs show the mean  $\pm$  SD ( $n = 3$  independent experiments) of the mitotic index at 30 hours

after transfection. (D) FRET assay comparing the MCC catalytic assembly (with 100 nM CDC20<sup>FL</sup>) with wild-type BUB1:BUB3 (light blue), BUB1<sup>ACM1</sup> (dark blue), BUB1<sup>KEN-ABBA</sup> (brown), or in the absence of BUB1:BUB3 (pink). (E) FRET assay comparing the MCC catalytic assembly with wild-type BUB1:BUB3 (light blue), BUB1<sup>mini</sup> (yellow), or in the absence of BUB1:BUB3 (pink). (F) FRET assay comparing the MCC catalytic assembly with wild-type catalysts (light blue), BUB1<sup>ACM1</sup> (dark blue), MAD1<sup>Q648A-R650A</sup> (dark green), CDC20<sup>BOX1-Glu</sup> (orange), or all mutants in combination (red).

as “catalysts”) come together at kinetochores to overcome the kinetic barrier and assemble the MCC (9, 10, 14, 15). We reconstituted catalytic activation of MCC assembly in vitro (9), but how the catalysts promote rapid MAD2:CDC20 association remains unknown and is a question of great interest with likely implications for other HORMA domain systems (16) (see supplementary text section 1).

Crucial for understanding this process is the precise order of events that precede assembly of the C-MAD2:CDC20 complex. Docking of O-MAD2 to the MAD1:C-MAD2 complex is an early step in catalytic assembly of MCC (9, 17, 18). It has been proposed that through this initial docking, O-MAD2 converts first into a ligand-free (empty) C-MAD2 intermediate, which then binds rapidly the CDC20 MIM (19, 20) (fig. S1A). When testing this idea with a combination of fluorescence-based in vitro sensors (figs. S1C, S2, and S3 and tables S1 and S2), however, we found that empty C-MAD2 did not associate to the MIM of full-length CDC20 (CDC20<sup>FL</sup>) at an appreciable rate (Fig. 1B and figs. S3H and, S4A). Binding of the MIM to empty C-MAD2 was only observed when the MIM sequence was presented to C-MAD2 as a peptide (CDC20<sup>MIM</sup>)

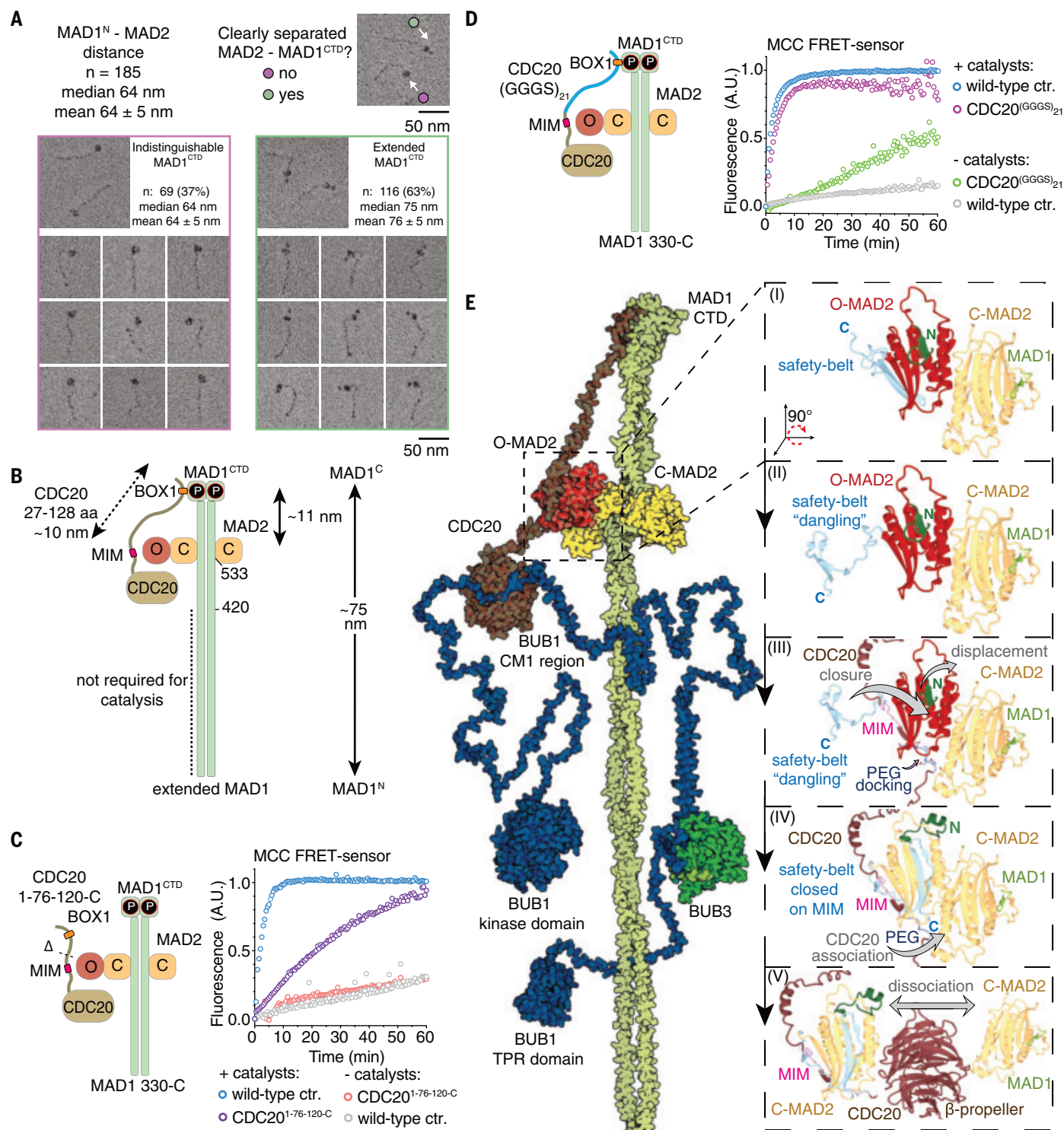
or at the beginning of an N-terminally truncated mutant (CDC20<sup>111-C</sup>; Fig. 1C and figs. S3G and S4B), likely because in both cases the MIM can thread under the latched safety belt of C-MAD2 (fig. S1B). This path, however, is not available to CDC20<sup>FL</sup> (Fig. 1B and fig. S3H) or to a C-terminal truncation mutant of CDC20 with substantial runs of residues on either side of the MIM (CDC20<sup>1-180</sup>) regardless of CDC20 phosphorylation (21, 22) (Fig. 1C and fig. S4, C to E). Furthermore, binding of O-MAD2 to CDC20<sup>MIM</sup> was not accelerated by catalysts beyond a very modest increase (10), whereas MCC assembly in the presence of CDC20<sup>FL</sup> was accelerated ~35-fold, suggesting that segments of CDC20 other than the MIM are also required for catalysis (Fig. 1D, fig. S5, and supplementary text section 2).

To determine the role of CDC20 in catalysis, we assessed the behavior of truncated CDC20 constructs (Fig. 2A and fig. S6, A and B). CDC20<sup>1-180</sup> associated spontaneously with O-MAD2, but the impairment of a crucial interaction of the CDC20 C-terminal  $\beta$ -propeller with BUB1:BUB3 (9) caused the reaction to be only moderately accelerated by catalysts (Fig. 2B). A very similar behavior was ob-

served with CDC20<sup>111-C</sup> (Fig. 2C), indicating that in addition to the CDC20  $\beta$ -propeller, the N-terminal extension is also required for efficient catalysis. The PEG motif (residues 138 to 144; fig. S6C) in the CDC20 N-terminal region is required for MAD2:CDC20 association but not for CDC20:APC/C association in vivo (3, 4, 23). A CDC20<sup>PEG</sup> mutant (CDC20<sup>PEG-Ala</sup>) associated with MAD2 and formed MCC but was partly insensitive to catalysts (Fig. 2D). Because a peptide encompassing the PEG motif in isolation did not accelerate the MAD2 conversion (fig. S3J), the PEG may contribute to docking the MIM of CDC20 onto MAD2 during catalysis.

The C-terminal region of MAD1 is required for SAC signaling (9, 24–26) and interacts with the N-terminal region of CDC20 through MPS1 phosphorylation (24). The BOX1 and BOX2 sequences in the CDC20 N-terminal region are evolutionarily conserved (fig. S6D). Without catalysts, BOX1 or BOX2 mutants (CDC20<sup>BOX1-Glu</sup> and CDC20<sup>BOX2-Glu</sup>) formed MCC-like wild-type CDC20; with catalysts, CDC20<sup>BOX1-Glu</sup>, but not CDC20<sup>BOX2-Glu</sup>, showed reduced rates of MCC assembly (Fig. 2E and fig. S6, E to G). Cross-linking mass-spectrometry (XL-MS) of the catalytic complex indicated





**Fig. 4. Spatial organization of the catalytic machinery promotes MIM accessibility.** (A) MAD1:MAD2 visualized by EM after glycerol spraying and low-angle platinum shadowing. The MAD1:MAD2 particles where the CTD is visible (green,  $n = 116$ ) or is not visible (pink,  $n = 69$ ) were manually picked and measured. Scale bars, 50 nm. (B) Schematic representation of the CDC20:MAD1 interaction. Reported measurements are based on available structural information. (C) FRET assay detecting the effect of CDC20<sup>1-76-120-C</sup> (purple) on the catalytic MCC assembly compared with CDC20<sup>FL</sup> (light blue). (D) FRET assay comparing the catalytic MCC assembly with CDC20<sup>FL</sup> (light blue)

or CDC20<sup>(GGGS)<sub>21</sub></sup> (violet). Note that CDC20<sup>(GGGS)<sub>21</sub></sup> was less stable than wild-type CDC20 and showed a tendency to precipitate over time. (E) Proposed float-and-catch model. Left, MAD1:C-MAD2 and BUB1:BUB3 complexes associate to facilitate the O-MAD2 and CDC20 interaction. Right, (I) O-MAD2 docks on C-MAD2; (II) the safety belt of O-MAD2 unfolds but the N terminus of O-MAD2 prevents its closure; (III) the PEG motif guides the MIM in position; (IV) CDC20 promotes the conformational change of MAD2 from O- to C-MAD2 by locking the safety belt on the MIM and displacing the N terminus of O-MAD2; and (V) C-MAD2 bound to CDC20 is released from MAD1:C-MAD2.

that BOX1 interacts with the C-terminal domain of MAD1 (MAD1<sup>CTD</sup>) (Fig. 3A and fig. S7). Docking and modeling identified a highly conserved antiparallel  $\beta$ -sheet of the RWD domain of MAD1 as a possible BOX1 receptor (fig. S8, A and B) (27). Mutations Q648A-R650A in MAD1 prevented CDC20 binding in vitro (fig. S8C) and caused a drop in the rate of MCC formation (Fig. 3B). When introduced in a MAD1-MIS12 chimera that enforces a permanent checkpoint arrest when expressed in HeLa cells (28), MAD1<sup>Q648A-R650A</sup> prevented enforcement of the SAC arrest (Fig. 3C). The extent of SAC inactivation in vitro and in vivo by MAD1<sup>Q648A-R650A</sup> was similar to the reduction previously observed with mutations of MPS1 phosphorylation sites on MAD1 (MAD1<sup>RWD-A</sup>; Fig. 3C and fig. S9) (9, 24). A triple alanine mutant of the conserved RLK motif on MAD1<sup>CTD</sup> (residues 617 to 619; MAD1<sup>RLK-AAA</sup>), which is required for SAC signaling in vivo (29), also failed to interact with CDC20 and strongly impaired catalytic MCC assembly (fig. S8, D and E). Systematic perturbation of functional motifs in the BUB1 N-terminal region revealed that only the conserved motif 1 (CM1), in addition to the CDC20 binding motifs (KEN, ABBA) (9, 30, 31), is required for catalysis in vitro (fig. S10, supplementary text section 3). In human cells, deletion of CM1 (BUB1<sup>ΔCM1</sup>) leads to partial loss of kinetochore MAD1 and an impaired SAC response (31, 32). Under standard assay conditions, BUB1<sup>ΔCM1</sup> retained substantial catalytic proficiency (fig. S10B), but more physiological concentrations of CDC20 (100 nM instead of 500 nM) exposed the defect of BUB1<sup>ΔCM1</sup> in MCC assembly, with effects comparable to omission of BUB1:BUB3 or its replacement with the BUB1<sup>KEN-ABBA</sup> mutant (Fig. 3D) (9, 33). Albeit less effectively than BUB1:BUB3, a minimal BUB1 construct encompassing the CM1 and KEN-ABBA sites (BUB1<sup>mini</sup>) sustained catalytic MCC formation (Fig. 3E), implying the formation of a catalytic intermediate in which CDC20 interacts simultaneously with MAD1 (through BOX1) and BUB1 (through KEN-ABBA) and the two catalysts further interact through CM1. In agreement with this hypothesis, when the MAD1<sup>Q648A-R650A</sup>, CDC20<sup>BOX1-Glu</sup>, and BUB1<sup>ΔCM1</sup> mutants were combined in a single reaction, catalysis was abolished (Fig. 3F and fig. S10, C to E). CDC20 promotes its own incorporation in the MCC as a substrate in cis, and not through a second CDC20 molecule associated with the catalyst (fig. S6, H to K, and supplementary text section 4).

By low-angle metal shadowing electron microscopy (EM), MAD1:MAD2 appears as an ~60-nm dimeric rod, with C-MAD2 defining two globular shapes at a distance of ~12 nm from the RWD “cap” in MAD1<sup>CTD</sup> (Fig. 4A) (11, 27, 34). This implies that the ~120 resi-

dues of CDC20 encompassing the BOX1 and the MIM motifs must extend for ~10 to 12 nm (Fig. 4B). Supporting this idea, the rate of MCC assembly was reduced after the distance between the BOX1 and MIM motifs (CDC20<sup>L76-120-C</sup>) was decreased by 44 residues (Fig. 4C). Conversely, after we replaced sequences upstream of BOX1 and between BOX1 and the MIM with highly flexible GGS repeats (CDC20<sup>(GGS)<sub>21</sub></sup>; table S2) without affecting the number of residues of CDC20, MCC assembly rates with catalysts were similar to those of wild-type CDC20 (Fig. 4D). Thus, the BOX1 and MIM motifs and a minimal distance between them, but not a specific conformation of the CDC20 N-terminal region, are required for catalysis.

There are contradictory reports on the binding site of BUB1 on MAD1 (33, 35, 36). Because of its low affinity, we were unable to confidently map this binding site. In *C. elegans*, the CM1 of BUB1 binds the coil-coiled region of MAD1 between residues 420 and 485 (35), a region also required for catalysis in vitro (fig. S8F and supplementary text section 3) (9). We propose that BUB1, bound to the CDC20  $\beta$ -propeller, docks distally to C-MAD2 relative to the MAD1<sup>CTD</sup>, promoting a “float-and-catch,” MPS1-phosphorylation-dependent interaction between CDC20<sup>BOX1</sup> and MAD1<sup>CTD</sup> (Fig. 4E). Thus, MAD1 “stretches” the N-terminal extension of CDC20 without external energy input to allow rapid access of O-MAD2 to the MIM, which may be otherwise inaccessible because of an intramolecular interaction with the C-terminal  $\beta$ -propeller (15) or a weakly folded conformation. After C-MAD2:CDC20 complex formation, BUB1:BUB3 may replace BUB1 on the  $\beta$ -propeller of CDC20, cementing MCC into a very stable complex (9).

Collectively, our results demonstrate that MAD2 and CDC20 must bind concomitantly to the catalytic complex for catalytic MCC assembly. This provides a plausible molecular explanation for why mitotic kinetochores, where MAD2, CDC20, and the catalytic complex reside, are essential for MCC assembly (1). In an accompanying paper, Lara-Gonzalez *et al.* report an orthogonal analysis of SAC signaling in *C. elegans*, and their conclusions are fully consistent with those described here (37).

## REFERENCES AND NOTES

1. A. Musacchio, *Curr. Biol.* **25**, R1002–R1018 (2015).
2. C. L. Rieder, R. W. Cole, A. Khodjakov, G. Sluder, *J. Cell Biol.* **130**, 941–948 (1995).
3. C. Alfieri *et al.*, *Nature* **536**, 431–436 (2016).
4. M. Yamaguchi *et al.*, *Mol. Cell* **63**, 593–607 (2016).
5. D. Izawa, J. Pines, *Nature* **517**, 631–634 (2015).
6. V. Sudakin, G. K. Chan, T. J. Yen, *J. Cell Biol.* **154**, 925–936 (2001).
7. L. Lad, S. Lichtsteiner, J. J. Hartman, K. W. Wood, R. Sakowicz, *Biochemistry* **48**, 9503–9515 (2009).

8. X. Luo *et al.*, *Nat. Struct. Mol. Biol.* **11**, 338–345 (2004).
9. A. C. Faesen *et al.*, *Nature* **542**, 498–502 (2017).
10. M. Simonetta *et al.*, *PLoS Biol.* **7**, e1000010 (2009).
11. L. Sironi *et al.*, *EMBO J.* **21**, 2496–2506 (2002).
12. X. Luo, Z. Tang, J. Rizo, H. Yu, *Mol. Cell* **9**, 59–71 (2002).
13. D. Izawa, J. Pines, *J. Cell Biol.* **199**, 27–37 (2012).
14. A. Kulukian, J. S. Han, D. W. Cleveland, *Dev. Cell* **16**, 105–117 (2009).
15. J. S. Han *et al.*, *Mol. Cell* **51**, 92–104 (2013).
16. S. C. Rosenberg, K. D. Corbett, *J. Cell Biol.* **211**, 745–755 (2015).
17. A. De Antoni *et al.*, *Curr. Biol.* **15**, 214–225 (2005).
18. M. Mapelli, L. Massimiliano, S. Santaguida, A. Musacchio, *Cell* **131**, 730–743 (2007).
19. X. Luo, H. Yu, *Structure* **16**, 1616–1625 (2008).
20. H. Yu, *J. Cell Biol.* **173**, 153–157 (2006).
21. T. Kim *et al.*, *Genes Dev.* **31**, 1089–1094 (2017).
22. E. Chung, R. H. Chen, *Nat. Cell Biol.* **5**, 748–753 (2003).
23. P. Lara-Gonzalez *et al.*, *Dev. Cell* **51**, 313–325.e10 (2019).
24. Z. Ji, H. Gao, L. Jia, B. Li, H. Yu, *eLife* **6**, e22513 (2017).
25. S. Heinrich *et al.*, *EMBO Rep.* **15**, 291–298 (2014).
26. T. Kruse *et al.*, *EMBO Rep.* **15**, 282–290 (2014).
27. S. Kim, H. Sun, D. R. Tomchick, H. Yu, X. Luo, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 6549–6554 (2012).
28. M. Maldonado, T. M. Kapoor, *Nat. Cell Biol.* **13**, 475–482 (2011).
29. D. M. Brady, K. G. Hardwick, *Curr. Biol.* **10**, 675–678 (2000).
30. B. Di Fiore *et al.*, *Dev. Cell* **32**, 358–372 (2015).
31. M. Vleugel *et al.*, *J. Cell Sci.* **128**, 2975–2982 (2015).
32. G. Zhang *et al.*, *EMBO J.* **38**, 4385–18 (2019).
33. G. Zhang *et al.*, *Nat. Commun.* **8**, 15822 (2017).
34. L. A. Allan *et al.*, *EMBO J.* **39**, e103180 (2020).
35. M. W. Moyle *et al.*, *J. Cell Biol.* **204**, 647–657 (2014).
36. N. London, S. Biggins, *Genes Dev.* **28**, 140–152 (2014).
37. P. Lara-Gonzalez, T. Kim, K. Oegema, K. D. Corbett, A. Desai, *Science* **371**, 10.1126/science.abc1424 (2020).

## ACKNOWLEDGMENTS

We thank all members of the Musacchio laboratory for helpful discussions; J. Deroissart, S. Carnignani, S. Sethi, C. Körner, S. Wohlgemuth, and H. Hausmann for help with reagents preparation; F. Müller, A. Brockmeyer, and P. Janning for mass spectrometry; and A. Desai and P. Lara-Gonzalez for fruitful discussions and for sharing unpublished data. **Funding:** A.M. gratefully acknowledges funding by the Max Planck Society, the European Research Council (ERC) Advanced Investigator Grant RECIPIANCE (proposal 669686), and the DFG's Collaborative Research Centre (CRC) 1093. V.P. acknowledges EMBO for the award of the EMBO long-term fellowship (ALTF 669-2017). **Author contributions:** V.P. and A.M. conceived the research, designed experiments, analyzed results, and wrote the manuscript. V.P., A.A., and P.S. subcloned constructs and expressed and purified recombinant proteins. V.P. performed in vitro experiments. S.M. and G.A.S. performed the cellular assays. P.J.H. performed the EM experiments. I.R.V. and V.P. performed the molecular docking and structural modeling. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions of the paper are available in the main text or the supplementary materials.

## SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/371/6524/67/suppl/DC1  
Materials and Methods  
Supplementary Text  
Figs. S1 to S10  
Tables S1 to S3  
References (38–57)  
Data S1 to S4  
MDAR Reproducibility Checklist

7 April 2020; accepted 18 November 2020  
10.1126/science.abc1152



## MEMBRANES

# Nanoscale control of internal inhomogeneity enhances water transport in desalination membranes

Tyler E. Culp<sup>1</sup>, Biswajit Khara<sup>2</sup>, Kaitlyn P. Brickey<sup>1</sup>, Michael Geitner<sup>1</sup>, Tawanda J. Zimudzi<sup>3</sup>, Jeffrey D. Wilbur<sup>4</sup>, Steven D. Jons<sup>4</sup>, Abhishek Roy<sup>5</sup>, Mou Paul<sup>6</sup>, Baskar Ganapathysubramanian<sup>2</sup>, Andrew L. Zydney<sup>1</sup>, Manish Kumar<sup>7\*</sup>, Enrique D. Gomez<sup>1,3,8\*</sup>

Biological membranes can achieve remarkably high permeabilities, while maintaining ideal selectivities, by relying on well-defined internal nanoscale structures in the form of membrane proteins. Here, we apply such design strategies to desalination membranes. A series of polyamide desalination membranes—which were synthesized in an industrial-scale manufacturing line and varied in processing conditions but retained similar chemical compositions—show increasing water permeability and active layer thickness with constant sodium chloride selectivity. Transmission electron microscopy measurements enabled us to determine nanoscale three-dimensional polyamide density maps and predict water permeability with zero adjustable parameters. Density fluctuations are detrimental to water transport, which makes systematic control over nanoscale polyamide inhomogeneity a key route to maximizing water permeability without sacrificing salt selectivity in desalination membranes.

Nearly 80% of worldwide fresh water is used for agriculture, livestock, and energy applications, which places substantial stress on existing water sources in both developed and developing countries (1, 2). Technologies such as membrane filtration, distillation, and ion exchange are extensively used to purify water (3, 4); nonetheless, the energy requirement to remove dissolved solutes, particularly salt, from water remains high.

Reverse osmosis (RO) (5) occupies a 66% share of the global desalination capacity and produces nearly 21 billion gallons of water per day (6). It is also playing an increasingly important role in recycling and recovering fresh water from wastewater and other waste streams for both human and industrial use (7, 8). Recent progress in RO membrane synthesis has yielded methodologies to manufacture highly permeable desalination membranes by controlling the internal morphology, thickness, and feed surface area of the fully aromatic polyamide (PA) active layer (9–12). It is not clear, however, how the resulting nanoscale PA morphology is linked to the performance observed in such membranes.

We describe a methodology to quantify the effect of three-dimensional (3D) nanoscale variations in polymer mass on water transport

within the PA active layer for a series of four RO membranes (PA1 to PA4). The PA films were synthesized by a conventional interfacial polymerization reaction between aqueous diamine and organic acid chloride solutions directly on a porous polysulfone support membrane (13, 14) conducted in a commercial pilot-scale manufacturing line (see materials and methods for synthesis details). The performance of the synthesized membranes was evaluated using cross-flow filtration (table S1). To isolate morphological influences on water transport properties, the differences in chemical composition between the membranes used were minimized, as previously described (13). Fourier transform infrared spectroscopy profiles (fig. S1) confirm a nearly constant carboxylic acid-to-amide ratio. Through a combination of electron tomography, energy-filtered transmission electron microscopy, and solution-diffusion simulations, we find that nanoscale variations in density are detrimental to water transport in these membranes and that controlling these density fluctuations is crucial to maximize performance in RO membranes (15, 16).

Transmission electron microscopy (TEM) has not been able to quantitatively link the PA microstructure with desalination performance (17–19). When imaging in scanning TEM (STEM) mode using a high-angle annular dark-field (HAADF) detector, images are formed by Z-contrast (20), where, for a single-component system, the pixel intensity is directly related to the sampled mass. By extension, for an isolated PA film, the pixel intensity of a HAADF-STEM image is a function of the sample thickness, density, and pixel size. To decouple PA thickness and density in the electron microscope, 3D reconstructions of the nanoscale PA morphology are necessary. We achieve this through HAADF-STEM tomography, where a tilt series is aligned to create a 3D model that describes

the nanoscale surface and internal PA morphology (details in the supplementary materials; Fig. 1, A and B; figs. S2 to S5; and movies S1 to S4). Quantification of 3D models reveals that the PA void fraction and surface area are consistent with analysis of similar commercial RO membranes (fig. S6) (21).

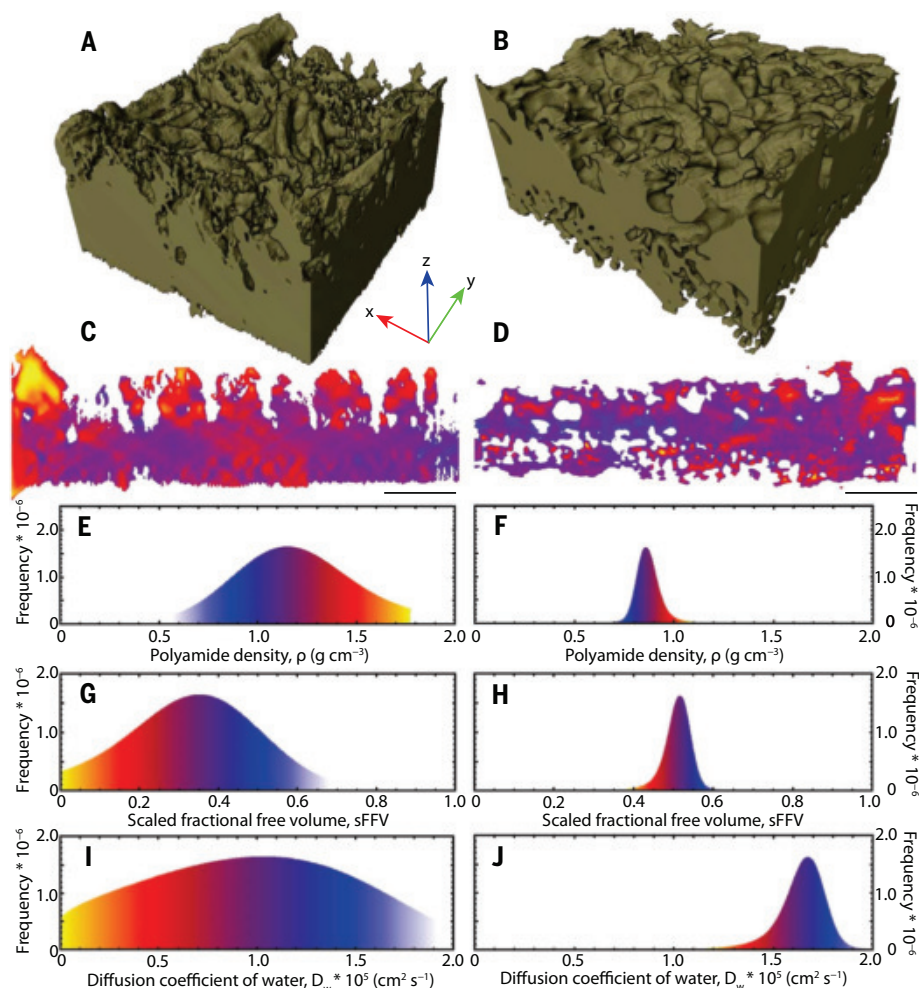
High-resolution HAADF-STEM tomography decouples PA density and thickness, which allows for the determination of nanoscale 3D distributions of each of these parameters independently. Although average values of relevant membrane properties are commonly used to estimate membrane transport rates, nanoscale distributions of mass likely govern transport through RO membranes (evident in Fig. 1, C and D). Variations in membrane resistance and water flux would therefore arise from a combination of nanoscale variations in PA thickness and density. We can use a combination of energy-filtered TEM and HAADF-STEM tomography to map how the nanoscale inhomogeneity in the density varies within PA films and how it relates to variations in membrane thickness. As described in detail in the supplementary materials, we convert the 3D nanoscale intensity distributions (Fig. 1, C and D) to nanoscale distributions of density ( $\rho$ ), from which we extract the water diffusivity ( $D_w$ ) within PA films (details in the supplementary materials; figs. S7 to S22; tables S2 and S3; and Fig. 1, E to J). We calculate the diffusivity from the density in a solution-diffusion formalism through the fractional free volume (FFV) by estimating the occupied volume from the maximum polymer density that we measured in our tomograms. This is effectively an excess free volume (22), where the occupied volume from the maximum density in 3D reconstructions is similar to reported values (23) and estimates from the Bondi method (24). We thus use the term scaled fractional free volume (sFFV) for values obtained using electron tomography to denote the introduction of this approach for estimating free volume. The color gradients under each curve for the density, sFFV, and water diffusivity shown in Fig. 1, E to J, serve as an absolute scale for the cross section in Fig. 1, C and D (figs. S19, S20, and S22 for PA2 and PA3).

A precursor for determining nanoscale 3D PA inhomogeneities is the measurement of the average values of PA density ( $\rho_{avg}$ ), the average sFFV ( $sFFV_{avg}$ ), and the average diffusion coefficient of water ( $D_{w,avg}$ ). In short, by accounting for the elastic and inelastic scattering components of a TEM image (Fig. 2, A to C), we can obtain the mean free path of electrons (details in the supplementary materials, figs. S7 to S11, and table S2). The mean free path can be used to determine  $\rho_{avg}$  and, by extension,  $sFFV_{avg}$  and  $D_{w,avg}$  in PA films (details in the supplementary materials; Fig. 2, D and E; and fig. S13). In the series of membranes we

<sup>1</sup>Department of Chemical Engineering, The Pennsylvania State University, University Park, PA 16802, USA. <sup>2</sup>Department of Mechanical Engineering, Iowa State University, Ames, IA 50011, USA. <sup>3</sup>Materials Research Institute, The Pennsylvania State University, University Park, PA 16802, USA. <sup>4</sup>DuPont Water Solutions, Edina, MN 55439, USA. <sup>5</sup>The Dow Chemical Company, Freeport, TX 77541, USA. <sup>6</sup>The Dow Chemical Company, Lake Jackson, TX 77566, USA. <sup>7</sup>Department of Civil, Architectural and Environmental Engineering, University of Texas, Austin, TX 78712, USA. <sup>8</sup>Department of Materials Science and Engineering, The Pennsylvania State University, University Park, PA 16802, USA.

\*Corresponding author. Email: manish.kumar@utexas.edu (M.K.); edg12@psu.edu (E.D.G.)





**Fig. 1. Quantifying the 3D nanoscale inhomogeneity of PA RO membranes through the combination of energy-filtered TEM and electron tomography.** (A and B) 3D isosurfaces of the PA1 (A) and PA4 (B) membranes. (C to J) 12-Å thick xz plane with colored voxels of PA1 (C) and PA4 (D) corresponding to colored gradients under the density [(E) and (F)], sFFV [(G) and (H)], and diffusion coefficient [(I) and (J)] of water histograms for the PA1 and PA4 membranes, respectively. (A), (C), (E), (G), and (I) show data for PA1; (B), (D), (F), (H), and (J) show data for PA4. All studied membranes show internal nanoscale inhomogeneity. Length axis arrows and scale bars are 200 nm. Histograms were obtained from reconstructions of PA films with  $>10^8$  voxels.

tested, the water permeance increased from  $6.39 \pm 0.22$  to  $8.36 \pm 0.15$  liters  $\text{m}^{-2}$   $\text{hour}^{-1}$   $\text{bar}^{-1}$  (LMH/bar) and was correlated to the  $\rho_{\text{avg}}$  decrease from  $1.15 \pm 0.14$  to  $0.86 \pm 0.09$   $\text{g cm}^{-3}$  (Fig. 2F). These average density values are in agreement with literature-reported bulk PA density values (25) and are consistent with an  $\sim 4$ - to  $4.6$ -Å spacing between chains, assuming liquid-like packing (with aligned chain backbones for simplicity). Further, over the same permeance increase,  $\text{sFFV}_{\text{avg}}$  increases from  $0.35 \pm 0.04$  to  $0.52 \pm 0.05$  (Fig. 2G), which indicates that increases in angstrom-scale free volume have a positive correlation with water flux. The large sFFV values are consistent with FFV predictions for glassy polymers and indicate that the PA free volume elements are likely interconnected (26).  $D_{\text{w,avg}}$  values were

obtained from a combination of free volume theory (27) and a compilation of  $D_{\text{w}}$  versus  $1/\text{sFFV}$  data from molecular dynamics simulations (fig. S21). The expected overall trend from the solution-diffusion model (26), commonly invoked to describe transport in RO membranes, was followed as  $D_{\text{w,avg}}$  increases from  $1.03 \pm 0.02$  to  $1.67 \pm 0.04 \times 10^{-5}$   $\text{cm}^2 \text{s}^{-1}$  and as the water permeance increases from  $6.39 \pm 0.22$  to  $8.36 \pm 0.15$  LMH/bar (Fig. 2H).  $D_{\text{w,avg}}$  values were consistent with values for PAs measured using quasi-elastic neutron scattering (15). These results indicate that  $\rho_{\text{avg}}$ ,  $\text{sFFV}_{\text{avg}}$ , and  $D_{\text{w,avg}}$  can be determined directly from TEM measurements.

The progressively narrowing  $D_{\text{w}}$  distributions from PA1 to PA4 as water permeance increases indicate that local distributions

in mass affect water transport. Nevertheless, water transport properties cannot be predicted exclusively from these distributions because the nanoscale water diffusivity distributions in Fig. 1, I and J, do not account for variations in membrane resistance. The spatial arrangement of localized membrane resistance variations plays a crucial role in determining diffusion pathways. Water molecules would be more likely to diffuse through a PA region of low thickness and density compared with a nearby thick and dense region—i.e., water transport would take the path of least resistance. Further, these variations in resistance would cause distributions in flow, causing flux hot spots (28), which cannot be accounted for purely on the basis of local or simply averaged  $D_{\text{w}}$ . The inability of average values to reliably predict transport properties is further highlighted by water flux calculations based on the average values reported in Fig. 2, E and H (i.e., for a PA film of uniform density and thickness), and the solution-diffusion model of water permeating through a nonporous membrane. The predicted water flux, based solely on average values, indicates decreasing water flux with increasing average thickness (table S4), which is opposite to the observed trend. The observed trend of increasing water flux with increasing thickness in this series of membranes (13) is counterintuitive and differs from previously reported results (29).

To predict transport properties, we calculate water diffusion through 3D models that show how thickness and  $D_{\text{w}}$  vary locally, which are obtained from the combination of energy-filtered TEM and electron tomography. We solve for the nanoscale variations in water transport by applying Fick's law with zero adjustable parameters at every  $1.7\text{-nm}^3$  voxel, totaling  $>100$  million voxels per 3D model. We ignore frame-of-reference effects (30) given that the water content is  $<15$  vol % (21, 31). This allows us to connect directly with simulations that estimate diffusion coefficients where frame-of-reference effects are also neglected (fig. S21). The boundary conditions are the concentrations of water at both the membrane feed and permeate surfaces determined by means of the solution-diffusion model (see calculation and details in the supplementary materials) (31). As a result, water diffusion pathways through PA films can be determined (Fig. 3) where the effect of nanoscale PA morphology on 3D water transport can be visualized. Light gray regions correspond to regions of ultralow water diffusivity within the membrane ( $D_{\text{w}} < 5 \times 10^{-6}$   $\text{cm}^2 \text{s}^{-1}$ ), which correspond to regions of high PA density and low sFFV. Dark gray regions correspond to water diffusivity between  $1.2$  and  $1.5 \times 10^{-5}$   $\text{cm}^2 \text{s}^{-1}$ . The regions of greatest resistance are near the PA top surface, which emphasizes the importance of PA surface area on water transport

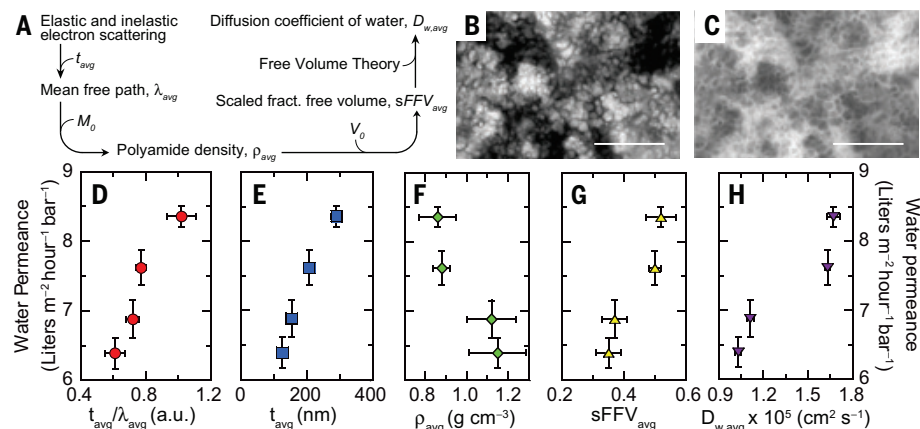
rates. Water diffusion pathways show variations in the  $x$  and  $y$  directions to avoid these areas, which indicates that water flow in the commonly seen surface polyps is low. Using the flow maps, we can reliably predict water permeability with zero adjustable parameters (Fig. 4).

The insets of Fig. 4 denote  $xy$  planes corresponding to predicted water flux distributions ( $J_{w,p}$ ) into and out of the reconstructed

PA volume as a result of density and thickness inhomogeneities for the PA1 and PA4 membranes (PA2 and PA3 are shown in figs. S23 and S24). Although all membranes show some local inhomogeneity in water flux, we find that the highest flux membrane (PA4) minimizes low-flow regions. A comparison of the flow distributions at several  $xy$  planes within the membrane and on each surface reveals evidence of lateral water transport ( $x$

and  $y$  directions), which indicates that the nanoscale PA morphology affects water transport in all three dimensions (Fig. 4, inset, and figs. S25 and S26). Using the calculated flow maps, we can reliably evaluate the predicted water permeability,  $P_{w,p}$ , showing qualitative agreement with measured water permeabilities from cross-flow filtration testing,  $P_{w,m}$  (Fig. 4). Small deviations between  $P_{w,p}$  and  $P_{w,m}$  could result from unaccounted effects introduced by the polysulfone support layer (9).  $P_{w,p}$  values are upward of 27% greater than water permeability predicted from a smooth PA film with a single homogeneous permeability (with the exception of PA3), which indicates that nanoscale internal inhomogeneities have a large effect on water transport in thick membranes (table S5).

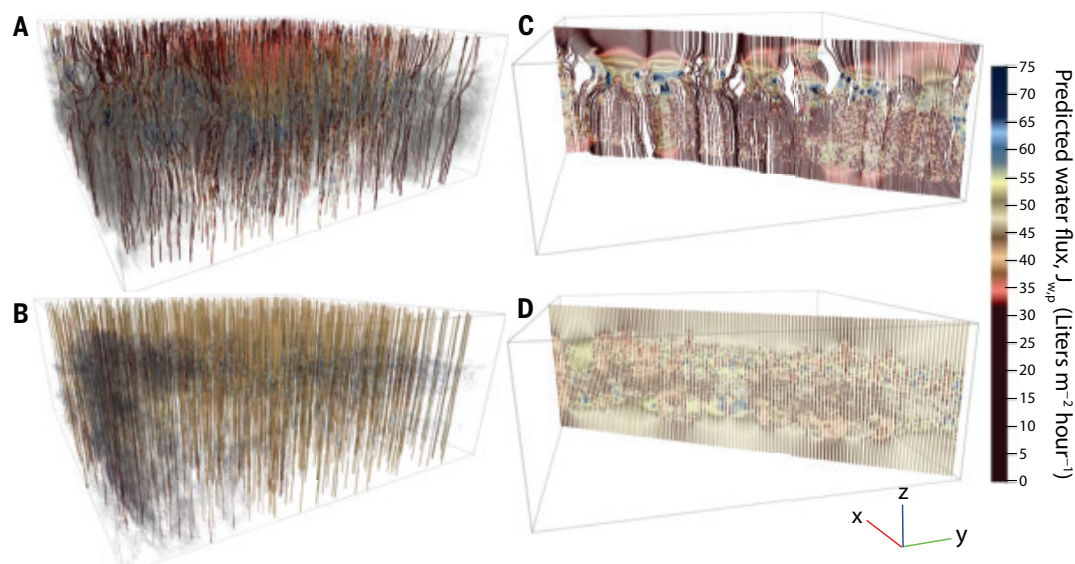
The most permeable membrane (PA4) has the lowest average density and narrowest density distribution, which suggests that highly permeable membranes minimize mass fluctuations that suppress water permeability, thereby maximizing overall permeability while maintaining selectivity. This is consistent with strategies to maximize both permeability and selectivity in gas separation membranes (32). Confining density values to a narrow distribution just below the boundary where the solute selectivity trade-off would be compromised could provide the highest possible water permeability for a desired selectivity. This would likely be in the form of a uniform density and resistance-selective region within the PA film, previously discovered to exist primarily near the PA feed surface (21) and illustrated by data shown for PA4 in Fig. 1 and density-thickness profiles in fig. S27. PA1 to PA3 have broader



**Fig. 2. Determining the average density, free volume, and diffusion coefficient of water in PA films using energy-filtered TEM.** (A) Flow chart of the process to determine the average diffusion coefficient of water in PA films. (B and C) Selected areas from a zero-loss (B) and composite thickness map image (C) of the PA4 membrane. Scale bars, 500 nm. (D and E) Water permeance as a function of average thickness map intensity,  $t_{\text{avg}}/\lambda_{\text{avg}}$ , (D) determined from (B) and (C), and average thickness,  $t_{\text{avg}}$  (E) (measured by means of ellipsometry). (F to H) Water permeance as a function of average PA density,  $\rho_{\text{avg}}$ , (F) determined from (D) and (E); PA monomer molecular weight,  $M_0$ , (G) sFFV<sub>avg</sub> determined from (F); and PA monomer specific occupied volume,  $V_0$ , and diffusion coefficient of water,  $D_{w,\text{avg}}$ , (H) determined from (G) and free volume theory for PA1 to PA4 membranes (details in the supplementary materials). Error bars are standard deviations with  $N = 5$  for PA1 to PA3 and  $N = 6$  for PA4. a.u., arbitrary units.

**Fig. 3. Calculating water transport through 3D models obtained from energy-filtered TEM and electron tomography.** (A to D) Perspective views [(A) and (B)] and cross sections [(C) and (D)] of the water diffusion pathways through the PA1 [(A) and (C)] and PA4 [(B) and (D)] membranes.

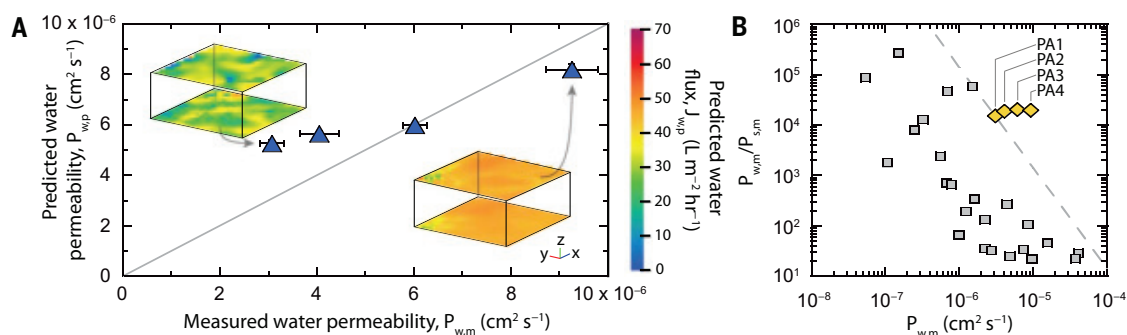
Gray areas in (A) represent regions of ultralow water diffusivity ( $D_w < 5 \times 10^{-6} \text{ cm}^2 \text{ s}^{-1}$ ) corresponding to regions of high PA density ( $\rho > 1.35 \text{ g cm}^{-3}$ ) and low sFFV (sFFV  $< 0.24$ ). Gray regions in (B) correspond to water diffusivity between  $1.2$  and  $1.5 \times 10^{-5} \text{ cm}^2 \text{ s}^{-1}$ . The streamlines are colored based on the local flux values in liters per square meter per hour. Larger regions of low water diffusivity in PA1 result in lateral water diffusion to avoid these high-resistance regions. Water diffusion pathways in PA4 show less lateral movement within the membrane, indicating smaller regions of high membrane resistance. Axis labels are scaled to 200 nm.





**Fig. 4. Nanoscale water transport calculations predict water permeability with zero adjustable parameters and comparison of membrane performance with state-of-the-art membranes.** (A) Predicted water flux ( $J_{w,p}$ ) maps detail the flow distribution for the PA1 and PA4 membranes (insets). By application of Fick's law, water permeability

can be predicted ( $P_{w,p}$ ) and compared with measured water permeability ( $P_{w,m}$ ). Flow distributions arise from nanoscale variations in PA resistance, where the PA1 flow distribution shows large inhomogeneities because of its broader density distribution. Conversely, the water flux distribution of PA4 is more homogeneous because of its narrower density distribution. Gray line serves as a guide to the eye. (B) Measured water-NaCl selectivity versus measured water permeability for desalination membranes used in this study compared with reported membranes. Yellow markers indicate PA1 to PA4 membranes and gray squares represent literature values. Desalination membrane data and upper bound line are from (31).



ity can be predicted ( $P_{w,p}$ ) and compared with measured water permeability ( $P_{w,m}$ ). Flow distributions arise from nanoscale variations in PA resistance, where the PA1 flow distribution shows large inhomogeneities because of its broader density distribution. Conversely, the water flux distribution of PA4 is more homogeneous because of its narrower density distribution. Gray line serves as a guide to the eye. (B) Measured water-NaCl selectivity versus measured water permeability for desalination membranes used in this study compared with reported membranes. Yellow markers indicate PA1 to PA4 membranes and gray squares represent literature values. Desalination membrane data and upper bound line are from (31).

density distributions in addition to higher average values and, thus, larger regions of low water diffusivity. We propose that a minimum average PA density exists that enables monovalent salt selectivity for suitable desalination specifications [e.g., a salt-to-water selectivity of 1:10,000 for seawater desalination (31)]. A membrane with an even narrower density distribution than PA4 approaching this threshold value would result in the effective water permeability approaching the true upper limit of fully aromatic PA RO membranes (for a given NaCl permeability). The synthesized RO membranes are near or above the upper-bound trade-off line for desalination membranes (Fig. 4B), which indicates that the open structure of PA4 already approaches this hypothesized upper limit for water permeability at brackish water salt retention levels (13).

We evaluated the generality of this approach with a similar analysis of a polyethersulfone ultrafiltration membrane used for virus filtration in downstream processing in the biopharmaceutical industry. 3D reconstructions reveal tortuous open pathways for water transport (fig. S28), and we again calculate flow properties. Despite different transport mechanisms compared with that of flow through dense PA films, an accurate effective diffusion coefficient and water flux can be obtained by accounting for the 3D pore network (see the supplementary materials).

The above methodology quantifies structure-property relationships for membranes that exceed literature-reported upper bounds of desalination performance, and it takes a step toward understanding water diffusion mechanisms and predicting transport rates. We demonstrate that the combination of energy-filtered TEM and electron tomography—i.e., multimodal electron microscopy—is a key tool to create predictive correlations between morphology and water transport for high-performance RO membranes. These cor-

relations can be extended to other molecular separation and polymeric systems to improve design strategies for various applications, including gas and hydrocarbon separations, carbon capture, blue energy production, and desalination.

#### REFERENCES AND NOTES

- M. A. Shannon *et al.*, *Nature* **452**, 301–310 (2008).
- C. J. Vörösmarty *et al.*, *Nature* **467**, 555–561 (2010).
- A. Deshmukh *et al.*, *Energy Environ. Sci.* **11**, 1177–1196 (2018).
- C. J. Johnson, P. C. Singer, *Water Res.* **38**, 3738–3750 (2004).
- C. Fritzmann, J. Löwenberg, T. Wintgens, T. Melin, *Desalination* **216**, 1–76 (2007).
- U. Caldera, C. Breyer, *Water Resour. Res.* **53**, 10523–10538 (2017).
- M. Qasim, M. Badrelzaman, N. N. Darwish, N. A. Darwish, N. Hilal, *Desalination* **459**, 59–104 (2019).
- S. B. Grant *et al.*, *Science* **337**, 681–686 (2012).
- M. R. Chowdhury, J. Steffes, B. D. Huey, J. R. McCutcheon, *Science* **361**, 682–686 (2018).
- S. Karan, Z. Jiang, A. G. Livingston, *Science* **348**, 1347–1351 (2015).
- Z. Tan, S. Chen, X. Peng, L. Zhang, C. Gao, *Science* **360**, 518–521 (2018).
- J. E. Gu *et al.*, *Adv. Mater.* **25**, 4778–4782 (2013).
- A. Roy *et al.*, Composite polyamide membrane having preferred azo content, U.S. Patent 9,555,378 (2017).
- R. J. Petersen, *J. Membr. Sci.* **83**, 81–150 (1993).
- E. P. Chan *et al.*, *Macromolecules* **53**, 1443–1450 (2020).
- P. M. Johnson, J. Yoon, J. Y. Kelly, J. A. Howarter, C. M. Stafford, *J. Polym. Sci. B Polym. Phys.* **50**, 168–173 (2012).
- F. A. Pacheco, I. Pinnau, M. Reinhard, J. O. Leckie, *J. Membr. Sci.* **358**, 51–59 (2010).
- Y. Li *et al.*, *J. Membr. Sci.* **534**, 9–17 (2017).
- T. E. Culp *et al.*, *ACS Macro Lett.* **7**, 927–932 (2018).
- C. Kübel *et al.*, *Microsc. Microanal.* **11**, 378–400 (2005).
- T. E. Culp *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 8694–8699 (2018).
- R. P. White, J. E. G. Lipson, *Macromolecules* **49**, 3987–4007 (2016).
- X. Zhang, D. G. Cahill, O. Coronell, B. J. Mariñas, *J. Membr. Sci.* **331**, 143–151 (2009).
- J. Park, D. Paul, *J. Membr. Sci.* **125**, 23–39 (1997).
- J. M. Dennison, X. Xie, C. J. Murphy, D. G. Cahill, *ACS Appl. Nano Mater.* **1**, 5008–5018 (2018).
- R. W. Baker, *Membrane Technology and Applications* (Wiley, 2007).
- J. Vrentas, J. Duda, *J. Polym. Sci. Polym. Phys. Ed.* **15**, 403–416 (1977).
- G. Z. Ramon, E. M. Hoek, *J. Membr. Sci.* **425–426**, 141–148 (2013).

- L. Lin, C. Feng, R. Lopez, O. Coronell, *J. Membr. Sci.* **498**, 167–179 (2016).
- D. R. Paul, *J. Membr. Sci.* **241**, 371–386 (2004).
- G. M. Geise, H. B. Park, A. C. Sagle, B. D. Freeman, J. E. McGrath, *J. Membr. Sci.* **369**, 130–138 (2011).
- H. B. Park, J. Kamcev, L. M. Robeson, M. Elimelech, B. D. Freeman, *Science* **356**, eaab0530 (2017).
- T. E. Culp *et al.*, Supporting data for “Nanoscale control of internal inhomogeneity enhances water transport in desalination membranes.” ScholarSphere (2020); <https://doi.org/10.26207/wk3w-wx50>.

#### ACKNOWLEDGMENTS

Flat-sheet membranes were provided courtesy of DuPont Water & Process Solutions in Edina, MN. The authors gratefully acknowledge educational discussions with B. Kuei and W. Song. The authors thank the instrumentation support available at the Materials Research Institute of The Pennsylvania State University and G. Foss at the Texas Advanced Computing Center (TACC) for help with visualization of simulation results. **Funding:** Financial support from The Dow Chemical Company and DuPont is acknowledged. T.E.C. and E.D.G. acknowledge financial support from the National Science Foundation under awards DMR-1609417 and DMR-1905550. K.P.B., A.L.Z., and E.D.G. also acknowledge support from the Center for Membrane Science, Engineering, and Technology (MAST) and the National Science Foundation under award IIP-1841474. M.K. acknowledges support from the National Science Foundation under award CBET-1946392. B.G. and B.K. are funded in part by the National Science Foundation under award CMMI-1906194. B.G. and B.K. also acknowledge computing support from XSEDE TG-CTS110007. **Author contributions:** J.D.W., S.D.J., A.R., M.P., B.G., A.L.Z., M.K., and E.D.G. designed the project. M.P. and A.R. designed and synthesized the membrane materials. T.E.C., K.P.B., M.G., and T.J.Z. performed sample preparation, data collection, and data analysis with assistance from M.K. and E.D.G. B.K. and B.G. performed nanoscale water transport calculations. J.D.W., S.D.J., A.R., M.P., B.G., A.L.Z., M.K., and E.D.G. supported in data analysis and interpretation. T.E.C., M.K., and E.D.G. wrote the paper with assistance from all coauthors. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are available in the manuscript, the supplementary materials, or at Penn State ScholarSphere (33).

#### SUPPLEMENTARY MATERIALS

[science.sciencemag.org/content/371/6524/72/suppl/DC1](https://science.sciencemag.org/content/371/6524/72/suppl/DC1)  
Materials and Methods  
Supplementary Text  
Figs. S1 to S28  
Tables S1 to S5  
References (34–54)  
Movies S1 to S4

23 March 2020; accepted 3 November 2020  
10.1126/science.abb8518



## MATERIALS SCIENCE

## Achieving large uniform tensile elasticity in microfabricated diamond

Chaoqun Dang<sup>1\*</sup>, Jyh-Pin Chou<sup>1,2\*</sup>, Bing Dai<sup>3\*</sup>, Chang-Ti Chou<sup>4\*</sup>, Yang Yang<sup>5</sup>, Rong Fan<sup>1</sup>, Weitong Lin<sup>1</sup>, Fanling Meng<sup>6</sup>, Alice Hu<sup>1,7†</sup>, Jiaqi Zhu<sup>3†</sup>, Jiecai Han<sup>3</sup>, Andrew M. Minor<sup>5</sup>, Ju Li<sup>8†</sup>, Yang Lu<sup>1,7,9†</sup>

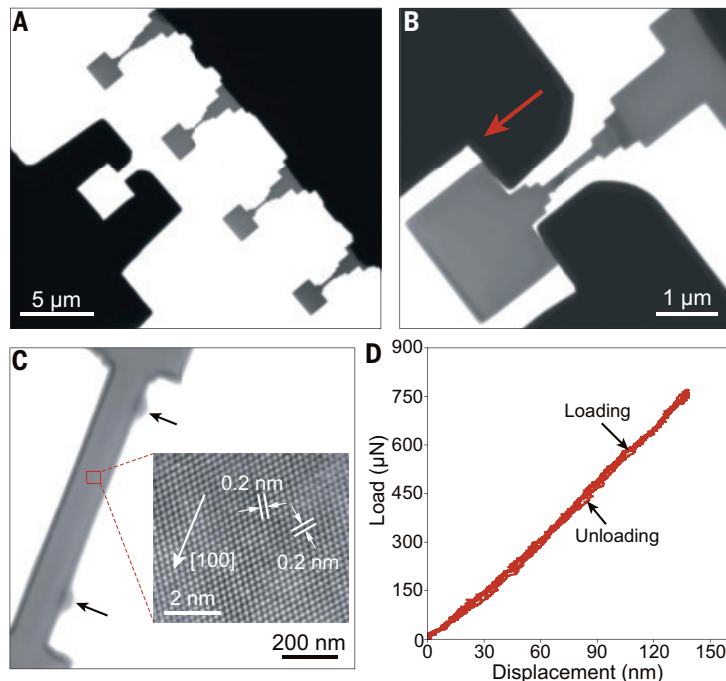
Diamond is not only the hardest material in nature, but is also an extreme electronic material with an ultrawide bandgap, exceptional carrier mobilities, and thermal conductivity. Straining diamond can push such extreme figures of merit for device applications. We microfabricated single-crystalline diamond bridge structures with ~1 micrometer length by ~100 nanometer width and achieved sample-wide uniform elastic strains under uniaxial tensile loading along the [100], [101], and [111] directions at room temperature. We also demonstrated deep elastic straining of diamond microbridge arrays. The ultralarge, highly controllable elastic strains can fundamentally change the bulk band structures of diamond, including a substantial calculated bandgap reduction as much as ~2 electron volts. Our demonstration highlights the immense application potential of deep elastic strain engineering for photonics, electronics, and quantum information technologies.

**D**iamond is the “Mount Everest” of electronic and photonic materials (1–3) because of its ultrahigh thermal conductivity (4), dielectric breakdown strength (2), carrier mobility (5), and ultrawide bandgap (6, 7). One serious obstacle to realizing diamond-based electronic and optoelectronic devices is the doping challenge (8–10) caused by the large bandgap and its crystal structure. A potential solution is to apply elastic lattice strain, which can substantially change the material properties (11–13). Ultralarge elastic deformation was recently demonstrated by bending a nanoscale diamond needle. The local tensile elastic strain reached 9% (14) or higher (15) in a region tens of nanometers in size, with the corresponding strength approaching diamond’s theoretical limit (16, 17). This discovery suggests that deep elastic strain engineering (ESE) (18, 19), in which very high (>5%) tensile and/or shear elastic strains are induced in diamond, may

allow for physical properties to be fundamentally changed. However, we require precise control in a large enough volume to fully use deep ESE for very large-scale integration in industry. Past attempts at straining diamond were often limited by the strain being within a small sample volume by flexural bending, resulting in nonuniform strain distribution (14, 15). Those samples are hard to control,

and the resulting high-strain field is highly localized. A large uniform elastic strain over substantial volume is often the desired initial state (18, 20) for deep ESE of an array of devices. This scenario is difficult to realize experimentally in a micrometer-scale sample, for example, in a clean wafer, because of the well-known “smaller is stronger” trend (12), which suggests that increasing size weakens the sample.

We demonstrate extremely large, reversible, and uniform elastic deformation in microfabricated single-crystalline diamond bridges under tensile loading. To produce tensile samples up to ~1-μm length by 300-nm width with well-defined geometry and crystal orientations, we used advanced microfabrication processes of bulk single-crystalline diamonds that were grown through microwave plasma-assisted chemical vapor deposition (21, 22). The process we developed produces high-quality diamond structures with micrometer dimensions, which are prime candidates for microelectromechanical systems (MEMS), quantum and photonic devices, arrays of strain-engineered transistors, and other applications. We used a homemade diamond tensile gripper to uniaxially stretch focused ion beam (FIB)-sculpted diamond from the bulk single crystal (23). We investigated key characteristics of



**Fig. 1. Microfabricated single-crystalline diamond bridge samples.** (A) Side-view TEM image showing the microfabricated diamond bridge samples and corresponding diamond tensile gripper. (B) Tensile sample and the diamond gripper aligned before straining. (C) Higher-magnification image depicting the gauge area with two fiducial markers (made by electron-beam-induced carbon deposition, as indicated by the black arrows), serving as “strain gauge” for the following strain measurement. Inset: HRTEM image depicting the atomic-scale structure of a pristine diamond tensile sample. (D) Typical load-versus-displacement curve read from the nanoindenter for a loading-unloading tensile test under the displacement control mode.

<sup>1</sup>Department of Mechanical Engineering, City University of Hong Kong, Kowloon, Hong Kong. <sup>2</sup>Department of Physics, National Changhua University of Education, Changhua 50007, Taiwan. <sup>3</sup>National Key Laboratory of Science and Technology on Advanced Composites in Special Environments, Harbin Institute of Technology, Harbin 150080, China. <sup>4</sup>Department of Materials Science and Engineering, National Chiao Tung University, Hsinchu 30010, Taiwan. <sup>5</sup>National Center for Electron Microscopy, Molecular Foundry, Lawrence Berkeley National Laboratory, and Department of Materials Science and Engineering, University of California, Berkeley, CA 94720, USA. <sup>6</sup>School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China. <sup>7</sup>Department of Materials Science and Engineering, City University of Hong Kong, Kowloon, Hong Kong. <sup>8</sup>Department of Nuclear Science and Engineering and Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. <sup>9</sup>Nano-Manufacturing Laboratory (NML), Shenzhen Research Institute of City University of Hong Kong, Shenzhen, 518057, China.

\*These authors contributed equally to this work.

†Corresponding author. Email: alicehu@cityu.edu.hk (A.H.); zhujq@hit.edu.cn (J.Z.); liju@mit.edu (J.L.); yanglu@cityu.edu.hk (Y.L.)

this reversible and uniform elastic deformation through in situ mechanical tensile experiments of the micrometer-sized diamond bridges at room temperature along the [100], [101], and [111] directions. We used density functional theory (DFT) calculations to estimate the electronic-band structure evolution upon the corresponding loading (24).

We used transmission electron microscopy (TEM) to characterize the microfabricated single-crystalline [100]-oriented diamond (25) (Fig. 1). Our low-magnification TEM image (Fig. 1A) shows several diamond tensile samples and the corresponding microfabricated diamond gripper (figs. S1 and S2). We FIB-sculpted the samples from the bulk diamond body (Fig. 1B), allowing the T-shaped sample to be gripped by the shoulders during the tensile test. We used two fiducial markers that we made using electron-beam-induced carbon deposition (Fig. 1C, black arrows) to serve as a “strain gauge.” Our high-resolution TEM (HRTEM) image shows the atomic-scale structure of a pristine diamond microtensile sample (Fig. 1C, inset). The surface of the FIB-sculpted diamond sample was an ~15-nm-thick amorphous carbon layer (fig. S3). This layer is commonly formed during FIB machining of the diamond (15, 17). A typical load-versus-displacement curve (Fig. 1D) that we measured using a quantitative nanoindenter

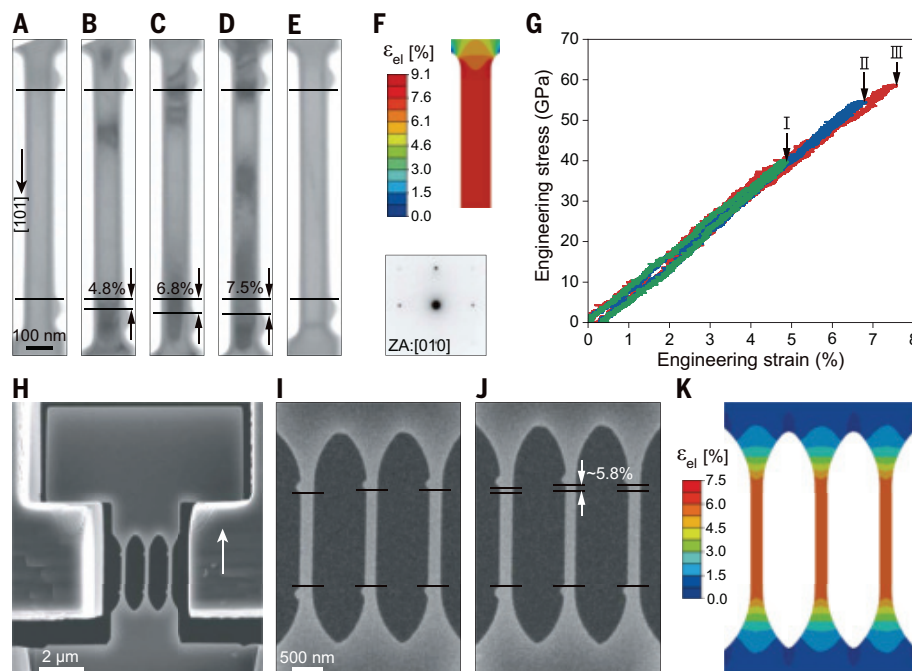
for a loading-unloading tensile test under the displacement control shows that fully elastic recovery was likely.

We tested >10 diamond samples for each crystallographic orientation inside a TEM (25) and videorecorded the sample evolution during straining. We extracted a series of TEM images (Fig. 2, A to E) from the video (movie S1) of in situ tension of a single-crystalline [101]-oriented diamond. We completed three complete loading-unloading processes with increasing tensile strain amplitude of a diamond bridge with an ~200-nm thickness. The diamond completely recovered its original length after strain values of ~4.8, 6.8, and 7.5% in each of these cycles, respectively. We used finite element method (FEM) simulation (Fig. 2F) (25) to reproduce the experimental setup (front view), which shows that the diamond has a highly uniform elastic strain distribution (~7.5%) in the interior, with a local maximum strain of ~9.1% near the gripping ends. The engineering stress-strain curves of loading-fully unloading tests indicate a nearly linear elastic response (Fig. 2G). The slope of the stress-strain curves gives a Young’s modulus of ~865 GPa, which is slightly lower than the theoretical Young’s modulus of bulk diamond along the [101] direction (~1100 GPa) (26), likely because of the surface amorphous carbon layer and growth imperfections (27).

We conducted a similar experiment for a diamond bridge along the [100] direction (25) (fig. S4 and movie S2).

To illustrate the potential for device applications, we further optimized the sample geometry using the ASTM standard (28) and microfabricated diamond array samples with multiple bridges (fig. S5C). We demonstrated in situ tensile straining of a diamond bridge array (length ~2  $\mu\text{m}$ ) in a scanning electron microscope (Fig. 2H and fig. S5) and show the loading-unloading process of a three-bridge array with increasing strain amplitude (movie S3). The diamond array completely recovered its original shape after being uniformly strained to ~5.8% (Fig. 2, I and J) and eventually fractured at ~6% (fig. S5E). We used an FEM simulation (Fig. 2K) to confirm the uniform elastic strain distribution (~5.8%) in the diamond array with minor stress concentrations near the gripping area.

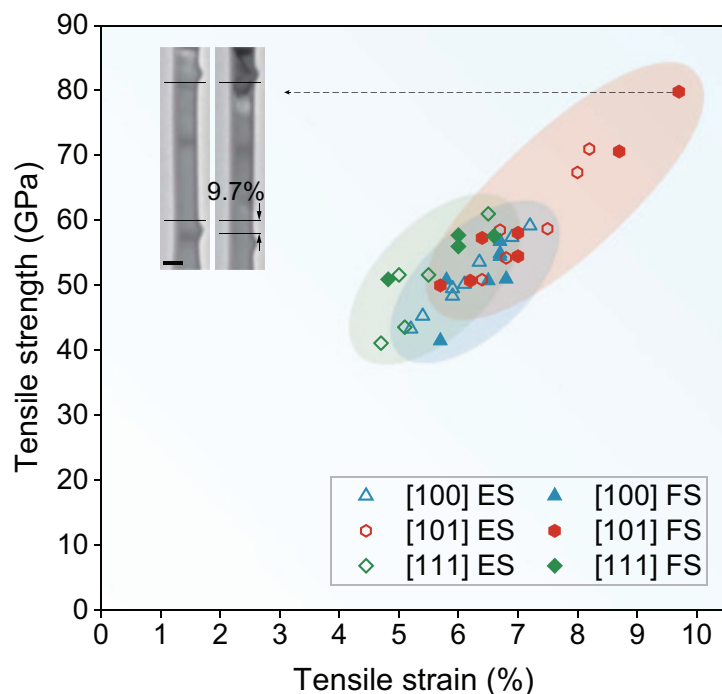
We compiled all the experimental data of the tensile strengths of [100]-, [101]-, and [111]-oriented diamond samples and plotted them against their tensile strains (Fig. 3) with their corresponding fracture morphologies (fig. S6) (25). The maximum elastic strains from the fully reversible runs and the failure run strains for each orientation are marked by the hollow and solid symbols, respectively. Loading-unloading experiments confirmed that the



**Fig. 2. Loading-unloading tensile experiments along the [101] direction.** (A to E) Elastic loading-unloading deformation of the diamond bridge sample with increased tensile straining amplitude and full unloading. The diamond recovered its original length after strain values of ~4.8, 6.8, and 7.5% in these cycles. (F) FEM simulation reproducing the critical geometry (front view) of the diamond bridge sample in (D) and showing the elastic strain ( $\epsilon_{el}$ ) distribution along the longitudinal direction. Inset: The [001] zone axis (ZA) selected area electron diffraction pattern depicting that the

loading is parallel to the [101] direction. (G) Corresponding engineering stress-versus-strain curves of multicycle loading-fully unloading tests (I, strain of 4.8%; II, strain of 6.8%; III, strain of 7.5%). (H) Tensile straining of a diamond array sample. (I to J) Elastic loading-unloading deformation of a [101]-oriented diamond array showing maximum 5.8% tensile strain with full recovery. (K) FEM simulation reproducing the critical geometry (front view) of the diamond array sample in (I) shows the highly uniform elastic strain ( $\epsilon_{el}$ ) distribution (~5.8%) along the tensile direction.





**Fig. 3. Statistical tensile results of [100]-, [101]-, and [111]-oriented diamonds.** Shown is a summary of the engineering tensile strengths of [100]-, [101]-, and [111]-oriented diamond bridge samples versus their tensile strains. The maximum elastic strains (ES) from fully reversible runs and the failure strains (FS) for each orientation are indicated by hollow and solid symbols, respectively. Inset shows a [101]-oriented diamond bridge sample with a maximum tensile strain of ~9.7%. Scale bar, 100 nm.

samples can consistently achieve above 6.5 to 8.2% sample-wide elastic strains with full recovery along the three different orientations. With the optimized sample geometry and microfabrication process, we can achieve maximum tensile strain up to 9.7% (Fig. 3, inset, and fig. S7). This value approaches the ideal elastic limit (16). The Young's modulus we measured is  $\sim 1010 \pm 70$  GPa in the [111] orientation, moderately higher than the values of  $\sim 895 \pm 65$  GPa and  $850 \pm 80$  GPa in the [101] and [100] orientations, respectively. These values are consistent with how the orientation-dependent elastic moduli rank in bulk diamond (26, 27).

As the experiments approach 10% uniform elastic strain, we performed DFT calculations from 0 to 12% strain (figs. S8 to S10 and table S1) to assess the impact on the electronic properties (fig. S11 and movie S4) (25). Generally, the bandgap of diamonds for each direction decreases as the tensile strain increases according to our simulation results. We used an electron energy-loss spectroscopy (EELS) analysis of a strained single-crystalline diamond sample (fig. S12) (25) to verify this trend. The simulation also shows that because the conduction band minimum position switches from the  $\Gamma$ - $X_2$  segment to the  $\Gamma$  point at ~9% strain acting along [111] direction (movie S4), strained diamond becomes a direct-bandgap semiconductor with

a bandgap of ~4.4 eV. Our DFT results indicate that the [101] direction has the largest bandgap reduction rate, down to 3.09 eV at 9% strain.

The fabrication of single-crystalline diamond bridge structures with micrometer-sized dimensions fits well with the scale of MEMS, photonic devices, quantum information processors, and arrays of microelectronic or nanoelectronic devices. The large and uniform elastic strains should drive changes in the bandgap, for which we found evidence using DFT simulation and EELS measurement. Straining along the [101] direction can induce a more substantial bandgap reduction compared with the other two directions. On the basis of our calculations, an indirect-direct bandgap transition may be possible with tensile strains larger than 9% along the [111] direction. These observations are an early step in potentially achieving deep ESE by microfabrication of the free-suspension array configurations and dynamic, reversible mechanical loading for diamond electronic, photonic, and quantum systems (29–31).

#### REFERENCES AND NOTES

1. P. W. May, *Science* **319**, 1490–1491 (2008).
2. C. J. H. Wort, R. S. Balmer, *Mater. Today* **11**, 22–28 (2008).
3. I. Aharonovich, A. D. Greentree, S. Praver, *Nat. Photonics* **5**, 397–405 (2011).
4. J. E. Field, *The Properties of Natural and Synthetic Diamond* (Academic, 1992).

5. J. Isberg *et al.*, *Science* **297**, 1670–1672 (2002).
6. H. Watanabe, C. E. Nebel, S. Shikata, *Science* **324**, 1425–1428 (2009).
7. J. Y. Tsao *et al.*, *Adv. Electron. Mater.* **4**, 1600501 (2018).
8. K. Okano, S. Koizumi, S. R. P. Silva, G. A. J. Amaratunga, *Nature* **381**, 140–141 (1996).
9. R. Kalish, *Diam. Relat. Mater.* **10**, 1749–1755 (2001).
10. Z. Teukam *et al.*, *Nat. Mater.* **2**, 482–486 (2003).
11. J. J. Gilman, *Electronic Basis of the Strength of Materials* (Cambridge Univ. Press, 2003).
12. T. Zhu, J. Li, *Prog. Mater. Sci.* **55**, 710–757 (2010).
13. J. Li, Z. Shan, E. Ma, *MRS Bull.* **39**, 108–114 (2014).
14. A. Banerjee *et al.*, *Science* **360**, 300–302 (2018).
15. A. Nie *et al.*, *Nat. Commun.* **10**, 5533 (2019).
16. D. Roundy, M. L. Cohen, *Phys. Rev. B* **64**, 212103 (2001).
17. J. M. Wheeler *et al.*, *Nano Lett.* **16**, 812–816 (2016).
18. Z. Shi *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **116**, 4117–4122 (2019).
19. C. Liu, X. Song, Q. Li, Y. Ma, C. Chen, *Phys. Rev. Lett.* **123**, 195504 (2019).
20. H. Zhang *et al.*, *Sci. Adv.* **2**, e1501382 (2016).
21. G. Shu *et al.*, *CrystEngComm* **20**, 198–203 (2018).
22. G. Shu *et al.*, *J. Cryst. Growth* **486**, 104–110 (2018).
23. D. Kiener, A. M. Minor, *Nano Lett.* **11**, 3816–3820 (2011).
24. W. Kohn, L. J. Sham, *Phys. Rev.* **140** (4A), A1133–A1138 (1965).
25. Materials and methods are available as supplementary materials.
26. R. H. Telling, C. J. Pickard, M. C. Payne, J. E. Field, *Phys. Rev. Lett.* **84**, 5160–5163 (2000).
27. J. E. Field, *Rep. Prog. Phys.* **75**, 126505 (2012).
28. ASTM Standard C1273-18, (ASTM International, 2018); <https://doi.org/10.1520/C1273-18>.
29. P. Ovarthaiyapong, K. W. Lee, B. A. Myers, A. C. B. Jayich, *Nat. Commun.* **5**, 4429 (2014).
30. B. Khanaliloo *et al.*, *Phys. Rev. X* **5**, 041051 (2015).
31. J. J. Hamlin, B. B. Zhou, *Science* **366**, 1312–1313 (2019).

#### ACKNOWLEDGMENTS

This work was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China, under project no. CityU1207416 and by the National Natural Science Foundation of China under project no. 11922215. J.Z. and B.D. are grateful for financial support from the National Science Fund for Distinguished Young Scholars (grant no. 51625201) and the National Natural Science Foundation of China (grant no. 51702066). A.H. acknowledges funding support from the City University of Hong Kong under project no. 9610336. Y.Y. and A.M.M. acknowledge support from the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division under contract no. DE-AC02-05-CH11231 to the Mechanical Behavior of Materials Program (KC13) at the Lawrence Berkeley National Laboratory (LBNL). Work at the Molecular Foundry was supported by the Office of Science, Office of Basic Energy Sciences, U.S. Department of Energy under contract no. DE-AC02-05CH11231. J.L. acknowledges support from Office of Naval Research MURI grant no. N00014-17-1-2661. **Author contributions:** Y.L. conceived the research. Y.L., A.H., J.Z., and J.L. supervised the research. C.D., J.-P.C., B.D., C.-T.C., Y.Y., and R.F. performed the research. C.D., J.-P.C., B.D., C.-T.C., Y.Y., R.F., W.L., A.M.M., J.L., and Y.L. analyzed the data. C.D., J.-P.C., B.D., C.-T.C., Y.Y., R.F., J.L., and Y.L. wrote the manuscript. All authors discussed the results and contributed to the final manuscript. **Competing interests:** The authors declare no competing financial interests. **Data and materials availability:** All data are reported in the paper or the supplementary materials.

#### SUPPLEMENTARY MATERIALS

[science.sciencemag.org/content/371/6524/76/suppl/DC1](https://science.sciencemag.org/content/371/6524/76/suppl/DC1)  
Materials and Methods  
Supplementary Text  
Figs. S1 to S12  
Table S1  
References (32–47)  
Movies S1 to S4

29 April 2020; accepted 23 November 2020  
10.1126/science.abc4174

## CORONAVIRUS

# Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Kenyan blood donors

Sophie Uyoga<sup>1\*</sup>†, Ifedayo M. O. Adetifa<sup>1,2†</sup>, Henry K. Karanja<sup>1†</sup>, James Nyagwange<sup>1†</sup>, James Tuju<sup>1</sup>, Perpetual Wanjiku<sup>1</sup>, Rashid Aman<sup>3</sup>, Mercy Mwangangi<sup>3</sup>, Patrick Amoth<sup>3</sup>, Kadondi Kasera<sup>3</sup>, Wangari Ng'ang'a<sup>4</sup>, Charles Rombo<sup>5</sup>, Christine Yegon<sup>5</sup>, Khamisi Kithi<sup>5</sup>, Elizabeth Odhiambo<sup>5</sup>, Thomas Rotich<sup>5</sup>, Irene Orgut<sup>5</sup>, Sammy Kihara<sup>5</sup>, Mark Otiende<sup>1</sup>, Christian Bottomley<sup>2</sup>, Zonia N. Mupe<sup>1</sup>, Eunice W. Kagucia<sup>1</sup>, Katherine E. Gallagher<sup>1,2</sup>, Anthony Etyang<sup>1</sup>, Shirine Voller<sup>1,2</sup>, John N. Gitonga<sup>1</sup>, Daisy Mugo<sup>1</sup>, Charles N. Agoti<sup>1</sup>, Edward Otieno<sup>1</sup>, Leonard Ndwiga<sup>1</sup>, Teresa Lambe<sup>6</sup>, Daniel Wright<sup>6</sup>, Edwine Barasa<sup>1</sup>, Benjamin Tsoka<sup>1</sup>, Philip Bejon<sup>1,6</sup>, Lynette I. Ochola-Oyier<sup>1</sup>, Ambrose Agweyu<sup>1†</sup>, J. Anthony G. Scott<sup>1,2†</sup>, George M. Warimwe<sup>1,6†</sup>

The spread of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in Africa is poorly described. The first case of SARS-CoV-2 in Kenya was reported on 12 March 2020, and an overwhelming number of cases and deaths were expected, but by 31 July 2020, there were only 20,636 cases and 341 deaths. However, the extent of SARS-CoV-2 exposure in the community remains unknown. We determined the prevalence of anti-SARS-CoV-2 immunoglobulin G among blood donors in Kenya in April–June 2020. Crude seroprevalence was 5.6% (174 of 3098). Population-weighted, test-performance-adjusted national seroprevalence was 4.3% (95% confidence interval, 2.9 to 5.8%) and was highest in urban counties Mombasa (8.0%), Nairobi (7.3%), and Kisumu (5.5%). SARS-CoV-2 exposure is more extensive than indicated by case-based surveillance, and these results will help guide the pandemic response in Kenya and across Africa.

**A**frica accounts for 17% of the global population (1) but by late July 2020 accounted for only 5% of the global COVID-19 cases and 3% of global COVID-19 deaths reported (2). This disparity has been attributed to limited capacity for diagnosis, timely implementation of stringent containment measures, a younger population structure, and a predominance of asymptomatic and mild infections (3, 4). The first case of COVID-19 in Kenya was detected on 12 March 2020. Within 1 week, the government instituted containment measures to limit the spread of the virus (5). By 31 July national surveillance recorded 20,636 cases and 341 deaths (6). This increase in cases is notably slower than the epidemic in Wuhan, Europe, or the United States. Recently, it has been suggested that “the virus is spreading... with an attenuated outcome in Africa” [(7), p. 626], but there are few data available to confirm or refute this assertion.

In countries affected early in the pandemic, serological surveillance was used to define cumulative incidence. For example, at the release of lockdown in Wuhan, 9.6% of staff resuming work were found to have antibodies

to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (8). At the end of the epidemic wave in Spain, seropositivity was 5.0% in a random population sample of 60,897 (9). As the epidemic curve declined in Geneva, seroprevalence rose over 3 weeks from 4.8 to 10.9% (10). Currently, there are few estimates of SARS-CoV-2 seroprevalence in Africa in the literature (11).

Movement restrictions, in response to COVID-19, have limited the conduct of fieldwork for population-based serosurveys. Several countries have monitored seroprevalence in blood transfusion donors (12, 13) or expectant mothers attending antenatal clinics (14). Here, we report the results of a pragmatic national serosurvey using residual blood samples from transfusion donors across Kenya and a highly sensitive and specific assay for anti-SARS-CoV-2 spike immunoglobulin G (IgG).

We validated a widely used enzyme-linked immunosorbent assay (ELISA) for SARS-CoV-2 IgG (15) with 910 serum samples from the pre-pandemic period and 174 sera from polymerase chain reaction (PCR)-defined SARS-CoV-2 cases, and a well-characterized five-sera panel from the National Institute of Biological Standards and Control (NIBSC) in the UK. For either receptor-binding domain (RBD) or whole spike, specificity was higher when using a ratio of the sample optical density (OD)/negative control OD than when using the raw sample OD plus 3 standard deviations to define seropositivity (table S1). By using OD ratios, both RBD and spike ELISAs correctly classified 901 of 910 prepandemic samples as seronegative (table S1). However, the spike ELISA

detected more seropositives (166 of 179 compared with 145 of 179 for RBD ELISA) among sera from SARS-CoV-2 PCR-positive individuals (fig. S2, A and B). On the basis of these data, we defined anti-SARS-CoV-2 IgG seropositivity as an OD ratio >2 and selected the spike ELISA for this study. The sensitivity and specificity, at this threshold, were 92.7% [95% confidence interval (CI), 87.9 to 96.1%] and 99.0% [95% CI, 98.1 to 99.5%], respectively (figs. S3, A and B, S5, and S6; and table S1). As previously noted (15), the RBD and whole-spike ELISA responses were highly correlated (fig. S3C), with very little interassay variation (fig. S4).

A total of 3174 blood transfusion samples were collected from four Kenya National Blood Transfusion Service (KNBTS) regional blood transfusion centers that are supported by several satellites and hospitals between 30 April and 16 June 2020, from individuals aged 15 to 66 years. Approximately half of the samples were drawn in Mombasa; the remainder were evenly distributed between Nairobi, Kisumu, and Eldoret (Fig. 1 and table S2). We excluded 18 duplicate samples, 56 records missing data on age or collection date, and two records from individuals aged ≥65 years. Policy in Kenya is to avoid blood donation from individuals >65 years, and we excluded these other data points as potentially unreliable. These exclusions left 3098 samples for further analysis (Fig. 1).

Of the 3098 samples, 174 were positive for anti-SARS-CoV-2 spike IgG, giving a crude seroprevalence of 5.6% (95% CI, 4.8 to 6.5%). Crude seroprevalence varied by age ( $P = 0.046$ ), ranging between 3.4 to 7.0% among adults 15 to 54 years; all 71 donors aged 55 to 64 years were seronegative (Table 1). Crude seroprevalence did not vary by sex ( $P = 0.50$ ) but did vary geographically, from 1.9% in the Rift Valley region to 10.0% in the Western region ( $P = 0.002$ ) (Table 1).

Compared with the 2019 Kenya Population and Housing Census, our participants were more commonly male (82.0% in our study versus 49.3% in the census), had more persons aged 25 to 34 years (40.1 versus 27.3%), and more residents of coastal counties (49.2 versus 9.1%) (Table 2). We therefore adjusted the prevalence estimate for the demographics of the sample using poststratification, and for the sensitivity and specificity of the test.

The Bayesian population-weighted and test-adjusted seroprevalence for Kenya was 4.3% (95% CI, 2.9 to 5.8%) (Table 1), and the posterior sensitivity and specificity estimates were 92.4% (95% CI, 88.0 to 95.6%) and 98.9% (95% CI, 98.2 to 99.5%), respectively. Seroprevalence was higher (4.2 to 5.2%) in the younger age groups (15 to 44 years) and declined in the older age groups (45 to 64 years) but was similar for both sexes. Seroprevalence was highest for those living in Mombasa, Nairobi, and the

<sup>1</sup>KEMRI-Wellcome Trust Research Programme, Kilifi, Kenya.

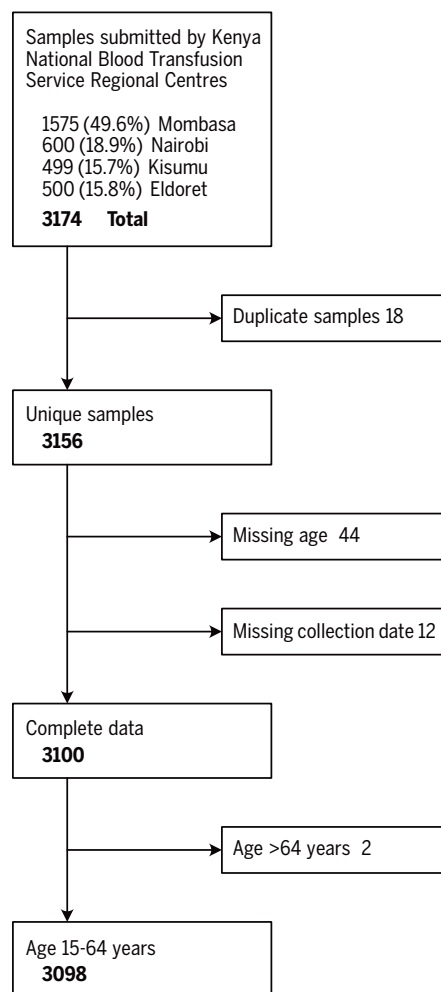
<sup>2</sup>Department of Infectious Diseases Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, London, UK.

<sup>3</sup>Ministry of Health, Government of Kenya, Nairobi, Kenya. <sup>4</sup>Presidential Policy and Strategy Unit, The Presidency, Government of Kenya, Nairobi, Kenya. <sup>5</sup>Kenya National Blood Transfusion Services, Ministry of Health, Nairobi, Kenya.

<sup>6</sup>Nuffield Department of Medicine, Oxford University, Oxford, UK.

\*Corresponding author. Email: suyoga@kemri-wellcome.org

†These authors contributed equally to this work. ‡These authors contributed equally to this work.

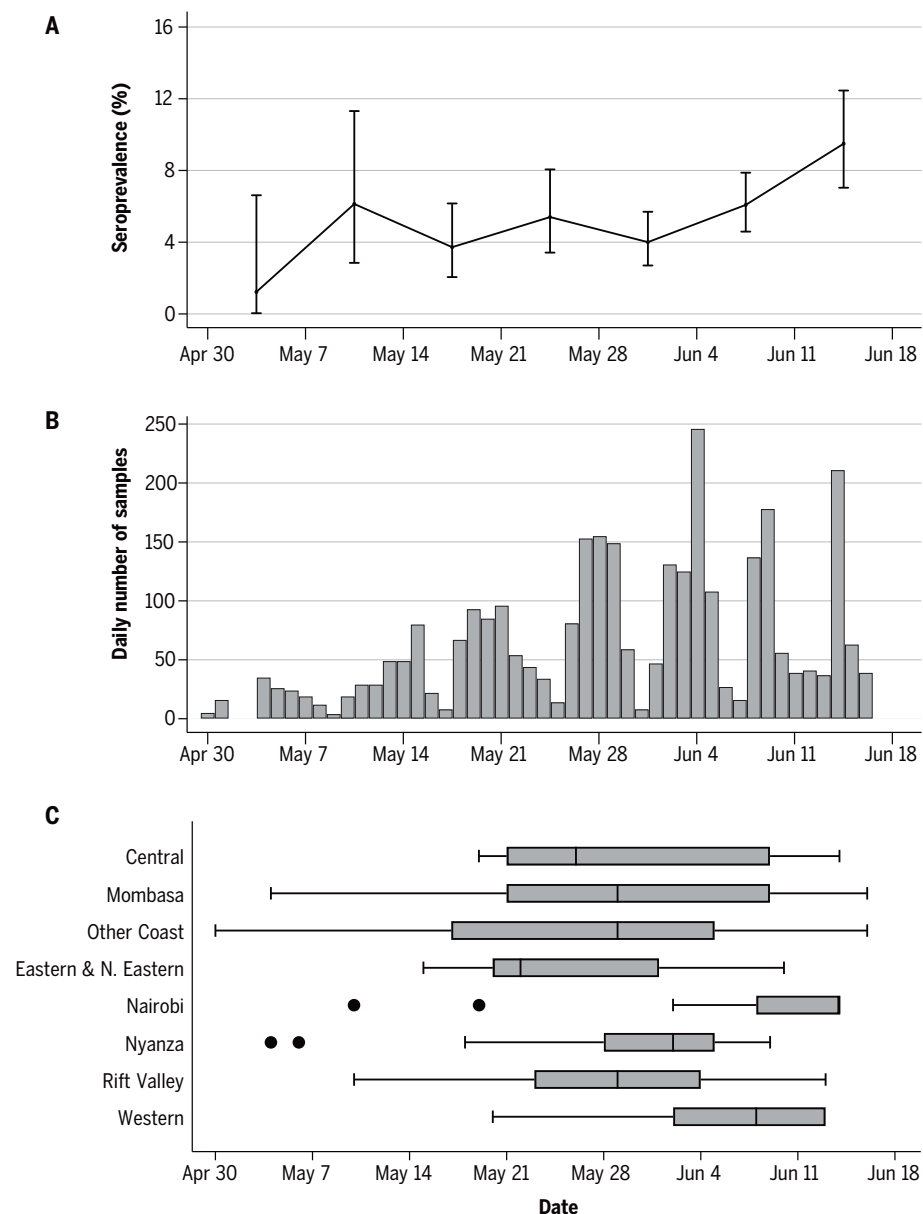


**Fig. 1. Participant flow diagram for SARS-CoV-2 seroprevalence study of blood donors in Kenya.** Exclusion criteria for the selection of samples with complete data.

Western region, although the number of observations for the Western region was small. The directly standardized seroprevalence estimates are presented in table S3. Seroprevalence was also calculated for counties that had at least 120 donors sampled. The three largest urban counties of Mombasa, Nairobi, and Kisumu had SARS-CoV-2 seroprevalence of 8.0% (95% CI, 5.5 to 11.1%), 7.3% (95% CI, 4.2 to 11.4%), and 5.5% (95% CI, 2.8 to 9.6%), respectively (table S4).

The frequency of blood donor sampling and crude seroprevalence estimates increased with time over the 7-week study period (Fig. 2). The median sample date was 30 May 2020, and the midpoint of the study was 24 May 2020. We did not adjust for sample date because the period of sampling varied for residents of different counties (Fig. 2C); instead, we show the variation in crude prevalence over time (Fig. 2A).

Voluntary nonremunerated donors (VNRDs), who donate blood at community-based “blood



**Fig. 2. Timeline of sampling for SARS-CoV-2 seroprevalence in blood donors in Kenya. (A to C)** Against the timeline of the sampling period, (A) is the weekly crude seroprevalence and 95% confidence interval, (B) is the daily frequency of samples collected, and (C) is the temporal distribution of samples by region. Shown are the proportion, counts, and regional distribution of donors during the study period.

drives,” made up only 7.6% (236 of 3098) of our sample of donors; the remainder were family replacement donors (FRDs) who provide a unit of blood in compensation for a transfusion received by a sick relative. The two groups did not differ significantly by age ( $P = 0.15$ ) or sex ( $P = 0.51$ ) (table S5). Crude seroprevalence was 8.5% (20 of 236) for VNRDs and 5.4% (154 of 2862) for FRDs. The median sample date for VNRDs (14 June 2020) was 2 weeks later than that for FRDs (29 May 2020).

Population exposure across Kenya, with a population-weighted test-adjusted seroprevalence of 4.3%, is considerably higher than

was previously thought, on the basis of the cases and deaths reported to date. Seroprevalence was particularly high in the three urban counties: Mombasa (8.0%), Nairobi (7.3%), and Kisumu (5.5%). Consistent with other studies, seroprevalence did not vary significantly by sex (9, 10, 16); however, it peaked in 35- to 44-year-olds and was lowest for those  $\geq 45$  years, which is also consistent with existing reports in which seroprevalence was found to be lower in older adults (9, 10).

SARS-CoV-2 seroprevalence in our study is comparable with estimates from large population-based serosurveys in China, Switzerland, Spain,



**Table 1. Crude, population-weighted, and test performance-adjusted SARS-CoV-2 anti-spike protein IgG seroprevalence by participant characteristics and regions.** Prevalence estimates were calculated by using multilevel regression and poststratification (MLRP) to account for differences in the sample population and the national population and subsequently adjusted for assay sensitivity and specificity.

	All samples	Seropositive samples	Crude seroprevalence		Kenya population (2019 Census)	Bayesian population-weighted seroprevalence*		Bayesian population-weighted, test-adjusted seroprevalence*	
			%	(95% CI)		%	(95% CI)	%	(95% CI)
Age									
15 to 24 years	808	49	6.1	4.5 to 7.9	9,733,174	5.1	3.7 to 6.9	4.4	2.7 to 6.4
25 to 34 years	1242	66	5.3	4.1 to 6.7	7,424,967	4.9	3.6 to 6.4	4.2	2.8 to 6.0
35 to 44 years	714	50	7.0	5.2 to 9.1	4,909,191	5.9	4.3 to 8.1	5.2	3.3 to 7.7
45 to 54 years	263	9	3.4	1.6 to 6.4	3,094,771	3.8	1.9 to 6.0	3.0	1.1 to 5.4
55 to 64 years	71	0	0		1,988,062	3.4	0.7 to 6.2	2.9	0.7 to 5.7
Sex									
Male	2540	146	5.7	4.9 to 6.7	13,388,243	4.4	2.9 to 6.2	3.6	1.9 to 5.8
Female	558	28	5.0	3.4 to 7.2	13,761,922	5.5	4.4 to 6.8	4.8	3.5 to 6.4
Regions									
Central	105	7	6.7	2.7 to 13.2	3,452,213	5.6	2.9 to 10.0	4.9	1.9 to 9.7
Mombasa	550	51	9.3	7.0 to 12.0	792,072	8.3	6.1 to 10.9	7.8	5.4 to 10.8
Other Coast	973	39	4.0	2.9 to 5.4	1,671,097	3.7	2.6 to 5.1	2.9	1.6 to 4.6
Eastern/N. Eastern	242	11	4.5	2.3 to 8.0	5,176,080	4.3	2.5 to 7.0	3.5	1.4 to 6.6
Nairobi	235	21	8.9	5.6 to 13.3	3,002,314	7.6	4.9 to 11.2	7.1	4.2 to 11.2
Nyanza	442	30	6.8	4.6 to 9.5	3,363,813	6.0	4.2 to 8.4	5.2	3.1 to 7.9
Rift Valley	481	8	1.7	0.7 to 3.3	7,035,581	2.1	1.1 to 3.6	1.5	0.4 to 3.1
Western	70	7	10.0	4.1 to 19.5	2,656,995	7.0	3.5 to 13.1	6.3	2.5 to 13.1
Total	3,098	174	5.6	4.8 to 6.5	27,150,165	4.9	3.9 to 6.2	4.3	2.9 to 5.8

\*Rewighted prevalence estimates based on demographic data from the 2019 Kenya Population and Housing Census

and the United States after the initial epidemic peak and after many tens of thousands of deaths (9, 10, 17, 18). Our results are also comparable with those of other surveys of blood donors in Brazil (13), Italy (12), and many parts of England (19). Kenya has an estimated population of 53 million in 2020, and 57% of the population is aged 15 to 64 years. If the transfusion donor seroprevalence of 4.3% was applied to all 15- to 64-year-olds, it would suggest approximately 1.3 million infections. However, by the median sample date, 30 May 2020, only 2093 cases had been detected (of which approximately 90% were asymptomatic), and there were 71 deaths among all ages (6). Although it is difficult to extrapolate our data directly to the whole population, they do strongly suggest that the infection is more widespread in Kenya than the current PCR test results suggest and indicate a need for more systematic testing. The current PCR testing strategy targets symptomatic individuals, health care workers, contacts of confirmed cases, international travelers, cross-border truck drivers, and residents of areas identified as “hotspots.”

What are the potential explanations for the divergence in the ratio of observed cases or deaths to serologically defined infections inferred from transfusion donors in Kenya, com-

pared with many high-income countries? (i) The seroprevalence could be overestimated because of bias in the selection or behavior of blood transfusion donors. (ii) Cases could be underascertained by national public health surveillance, although it seems unlikely that reporting of deaths and severe cases could be reduced by several orders of magnitude, and hospitals in Kenya were not overwhelmed by admissions with respiratory illness. (iii) The steep demographic age-pyramid results in a smaller vulnerable age group. In Kenya, only 3.9% of the population is aged 65 years or greater, which is substantially less than, for example, 23.3% found in Italy; again, this would only explain a moderate reduction in severe cases or deaths (4). (iv) There may be alternative mechanisms of immunity to SARS-CoV-2, including cell-mediated immunity (20, 21), perhaps as a result of human coronavirus (HCoV)-elicited immunity (22, 23). Despite our prior work showing that HCoVs circulate in Kenya (24), we did not identify evidence of cross-reactive antibodies to endemic coronaviruses in our validation study.

Although blood donors are not representative of the Kenyan population as a whole, we adjusted for demographic bias in the sample structure by standardization against the age, sex, and regional distribution of the Kenyan

population. A substantial proportion (43%) of the population of Kenya is outside the age range (15 to 64 years) sampled in this study, and the seroprevalence in children <15 years and adults >65 years is often lower (9, 10); our estimate for blood donors may be higher than the estimate for the population as a whole. Blood donors also differ from the general population in their risk of exposure to SARS-CoV-2. For example, potential donors are excluded from giving blood if they have been ill during the past 6 months, so the sample may underestimate the population prevalence of SARS-CoV-2 antibodies; however, people who are shielding at home are unlikely to be captured in our sample, leading to an overestimate of seroprevalence. Our exploration of the two distinct populations of blood donors, FRDs and VNRDs, suggests variation in the seroprevalence by donor group, but 92% (2862 of 3098) of our sample came from the group with lower seroprevalence, and exclusion of VNRDs reduced little the crude seroprevalence in our study, from 5.6 to 5.4%. Against these considerations, other countries have relied on blood transfusion donors for an early estimate of seroprevalence, but later estimates from random population samples have not been substantially different (25, 26).

**Table 2. General characteristics of the study population compared with the national population of Kenya.** *N* is the number of individuals in each stratum.

		Blood transfusion samples		Kenya National Census 2019	
		<i>N</i>	%	<i>N</i>	%
Age	15 to 24 years	808	26.1	9,733,174	35.8
	25 to 34 years	1,242	40.1	7,424,967	27.3
	35 to 44 years	714	23.0	4,909,191	18.1
	45 to 54 years	263	8.5	3,094,771	11.4
	55 to 64 years	71	2.3	1,988,062	7.3
Sex	Male	2540	82.0	13,388,243	49.3
	Female	558	18.0	13,761,922	50.7
Regions	Central	105	3.4	3,452,213	12.7
	Mombasa	550	17.8	792,072	2.9
	Other Coast	973	31.4	1,671,097	6.2
	Eastern/Northeastern	242	7.8	5,176,080	19.1
	Nairobi	235	7.6	3,002,314	11.1
	Nyanza	442	14.3	3,363,813	12.4
	Rift Valley	481	15.5	7,035,581	25.9
	Western	70	2.3	2,656,995	9.8
Total	Kenya 15 to 64 years	3098		27,150,165	

A key strength of this study is the rigorous validation that included testing positive and negative control samples from the target population, as well as reference plasma from the UK NIBSC as part of a World Health Organization (WHO)-coordinated effort on SARS-CoV-2 sero-epidemiology. In addition, we adopted a conservative seropositivity threshold to optimize assay specificity and sensitivity for our setting.

The pandemic response in countries with limited health care capacity has been driven by the aggressive implementation of control measures to limit transmission. Unfortunately, this strategy has been accompanied by enormous collateral costs, particularly in Africa. Modeled estimates of the disruptions of essential medical services, such as immunization and antenatal care, suggest an additional ~253,500 child deaths and 12,200 maternal deaths over 6 months in low- and middle-income countries (27). In the absence of social protection, the economic effects of lockdown are debilitating, so it is important to obtain an early measure of the trajectory of the epidemic.

Our study provides a national and regional estimate of population exposure to SARS-CoV-2 in an African country. The 4.3% prevalence in blood transfusion donors is in sharp contrast with the reported COVID-19 cases and deaths and supports the impression that disease may be attenuated in Africa (7).

REFERENCES AND NOTES

1. United Nations Department of Economic and Social Affairs Population Division, "World Urbanization Prospects: The 2018 Revision," custom data acquired via (2018); <https://population.un.org/wup/DataQuery>.

2. Africa Centres for Diseases Control and Prevention, "Coronavirus Disease 2019 (COVID-19)" (2020); <https://africacdc.org/covid-19> [accessed 21 July 2020].

3. B. Z. Diop, M. Ngom, C. Pougué Biyong, J. N. Pougué Biyong, *BMJ Glob. Health* **5**, e002699 (2020).

4. J. B. Dowd *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9696–9698 (2020).

5. Ministry of Health Kenya, "Press Statement on the update of the coronavirus in the country and response measures" (2020); [www.health.go.ke/wp-content/uploads/2020/03/Coronavirus-Press-Statement-March-17-2020.pdf](http://www.health.go.ke/wp-content/uploads/2020/03/Coronavirus-Press-Statement-March-17-2020.pdf).

6. Ministry of Health Kenya, "COVID-19 Situation Reports (SITREP)" (2020); [www.health.go.ke/#1591180376422-52af4c1e-256b](http://www.health.go.ke/#1591180376422-52af4c1e-256b).

7. M. Mbow *et al.*, *Science* **369**, 624–626 (2020).

8. X. Wu, B. Fu, L. Chen, Y. Feng, *J. Med. Virol.* **92**, 1795–1796 (2020).

9. M. Pollán *et al.*, *Lancet* **396**, 535–544 (2020).

10. S. Stringhini *et al.*, *Lancet* **396**, 313–319 (2020).

11. M. G. Chibwana *et al.*, *Wellcome Open Res.* **5**, 199 (2020).

12. L. Valenti *et al.*, SARS-CoV-2 seroprevalence trends in healthy blood donors during the COVID-19 Milan outbreak. *medRxiv* 2020.2005.2011.20098442 [Preprint] 31 May 2020.

13. L. Amorim Filho *et al.*, *Rev. Saude Publica* **54**, 69 (2020).

14. D. D. Flannery *et al.*, *Sci. Immunol.* **5**, eabd5709 (2020).

15. F. Amanat *et al.*, *Nat. Med.* **26**, 1033–1036 (2020).

16. A. T. Huang *et al.*, *Nat. Commun.* **11**, 4704 (2020).

17. X. Xu *et al.*, *Nat. Med.* **26**, 1193–1195 (2020).

18. F. P. Havers *et al.*, Seroprevalence of antibodies to SARS-CoV-2 in 10 sites in the United States, March 23–May 12, 2020. *JAMA Intern Med.* 10.1001/jamainternmed.2020.4130 (2020).

19. Public Health England, "Sero-prevalence epidemiology, England, in Weekly Coronavirus Disease 2019 (COVID-19) Surveillance Report 2020 Week 28" (2020); [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/899301/Weekly\\_COVID19\\_Surveillance\\_Report\\_week\\_28.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/899301/Weekly_COVID19_Surveillance_Report_week_28.pdf).

20. A. Grifoni *et al.*, *Cell* **181**, 1489–1501.e15 (2020).

21. N. Le Bert *et al.*, *Nature* **584**, 457–462 (2020).

22. A. Sette, S. Crotty, *Nat. Rev. Immunol.* **20**, 457–458 (2020).

23. K. W. Ng *et al.*, Pre-existing and de novo humoral immunity to SARS-CoV-2 in humans. *bioRxiv* 2020.2005.2014.095414 [Preprint] 23 July 2020.

24. G. P. Otieno *et al.*, *Wellcome Open Res.* **5**, 150 (2020).

25. H. Ward *et al.*, Antibody prevalence for SARS-CoV-2 following the peak of the pandemic in England: REACT2 study in 100,000 adults. *medRxiv* 2020.2008.2012.20173690/20173692 [Preprint] 21 August 2020.

26. Public Health England, National COVID-19 surveillance report week 40 (2020); [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/923668/Weekly\\_COVID19\\_Surveillance\\_Report\\_week\\_40.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/923668/Weekly_COVID19_Surveillance_Report_week_40.pdf).

27. T. Robertson *et al.*, *Lancet Glob. Health* **8**, e901–e908 (2020).

28. S. Uyoga *et al.*, Replication Data for: Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Kenyan blood donors. Harvard Database (2020); <https://doi.org/10.7910/DVN/RENV9C>.

ACKNOWLEDGMENTS

We thank F. Krammer for providing the plasmids to generate the RBD, spike protein, and CR3022 monoclonal antibody used in this work. Development of SARS-CoV-2 reagents was partially supported by the NIAID Centres of Excellence for Influenza Research and Surveillance (CEIRS) contract HHSN272201400008C. The COVID-19 convalescent plasma panel (NIBSC 20/118) and research reagent for SARS-CoV-2 Ab (NIBSC 20/130) were obtained from the NIBSC, UK. We thank the blood donors and KNBTS staff who supported this work. We also thank the WHO SOLIDARITY II network for sharing of protocols and for facilitating the development and distribution of control reagents. This paper has been published with the permission of the director, Kenya Medical Research Institute. This study was approved by the Scientific and Ethics Review Unit (SERU) of the Kenya Medical Research Institute (Protocol SSC 3426). Before the blood draw, donors gave individual consent for the use of their samples for research. Ethical approval was obtained for collection, storage, and further use for the sample sets used in the validation assays (SERU numbers 1433, 3149, and 3426). **Funding:** This project was funded by the Wellcome Trust (grants 220991/Z/20/Z and 203077/Z/16/Z), the Bill and Melinda Gates Foundation (INV-017547), and the Foreign Commonwealth and Development Office (FCDO) through the East Africa Research Fund (EARF/ITT/039). S.U. is funded by DELTAS Africa Initiative [DEL-15-003], L.I.O.-O. is funded by a Wellcome Trust Intermediate Fellowship (107568/Z/15/Z), A.A. is funded by a DFID/MRC/NIHR/Wellcome Trust Joint Global Health Trials Award (MR/R006083/1), J.A.G.S. is funded by a Wellcome Trust Senior Research Fellowship (214320) and the NIHR Health Protection Research Unit in Immunisation, I.M.O.A. is funded by the United Kingdom's Medical Research Council and Department For International Development through an African Research Leader Fellowship (MR/S005293/1) and by the NIHR-MPRU at UCL (grant 2268427 LSHTM). G.M.W. is supported by a fellowship from the Oak Foundation. C.N.A. is funded by the DELTAS Africa Initiative [DEL-15-003], and the Department for International Development and Wellcome (220985/Z/20/Z). **Author contributions:** Conceptualization and methodology: S.U., I.M.O.A., A.A., J.A.G.S., E.W.K., K.E.G., A.E., S.V., R.A., M.M., P.A., K.Ka., W.N., and G.M.W. Investigation: S.U., H.K.K., J.N., J.T., P.W., C.R., C.Y., K.Ki., E.Od., T.R., I.O., S.K., Z.N.M., J.N.G., D.M., C.N.A., E.Ot., L.N., L.I.O.-O., and G.M.W. Formal analysis: J.A.G.S., M.O., and C.B. Validation: T.L., D.W., H.K.K., J.N., J.T., L.I.O.-O., and G.M.W. Resources and funding acquisition: S.U., T.L., P.B., and G.M.W. Supervision: E.B., B.T., and P.B. Writing, original draft preparation: S.U., I.M.O.A., A.A., J.A.G.S., and G.M.W. Writing, review and editing: all authors. **Competing interests:** R.A., M.M., K.K., and P.A. are from the Ministry of Health, Government of Kenya. All other authors declare no competing interests. **Data and materials availability:** Deidentified data has been published on the Havard dataverse server (28). This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>. This license does not apply to figures/photos/artwork or other content included in the article that is credited to a third party; obtain authorization from the rights holder before using such material."

SUPPLEMENTARY MATERIALS

[science.sciencemag.org/content/371/6524/79/suppl/DC1](https://science.sciencemag.org/content/371/6524/79/suppl/DC1)  
Materials and Methods  
Figs. S1 to S6  
Tables S1 to S5  
References (29–32)  
MDAR Reproducibility Checklist

20 August 2020; accepted 6 November 2020  
Published online 11 November 2020  
10.1126/science.abe1916

## CLOUD PHYSICS

# Aerosol invigoration of atmospheric convection through increases in humidity

Tristan H. Abbott\* and Timothy W. Cronin

Cloud-aerosol interactions remain a major obstacle to understanding climate and severe weather. Observations suggest that aerosols enhance tropical thunderstorm activity; past research, motivated by the importance of understanding aerosol impacts on clouds, has proposed several mechanisms that could explain that observed link. We find that high-resolution atmospheric simulations can reproduce the observed link between aerosols and convection. However, we also show that previously proposed mechanisms are unable to explain the invigoration. Examining underlying processes reveals that, in our simulations, high aerosol concentrations increase environmental humidity by producing clouds that mix more condensed water into the surrounding air. In turn, higher humidity favors large-scale ascent and stronger convection. Our results provide a physical reason to expect invigorated thunderstorms in high-aerosol regions of the tropics.

Observations suggest that cloud-aerosol interactions play an important role in setting the frequency and intensity of atmospheric deep convection. Many studies have found increases in cloud top height and cloud cover coincident with increases in aerosol loading (1–5); additionally, lightning flash rates are consistently higher in high-aerosol regions of the tropics, including continents, islands, and ship tracks (6–8). These observations indicate that high aerosol concentrations may trigger a chain of processes that ultimately increases the number or strength of convective updrafts—which we refer to throughout this paper as “microphysical invigoration.” A deeper understanding of microphysical invigoration could enhance understanding of severe weather risks and of climatologically important cloud-aerosol interactions (9–12); this is particularly relevant because human activity is a major aerosol source.

Past work has proposed two mechanisms by which aerosol concentrations could invigorate convection. The first mechanism relies on a “cold-phase” pathway: Higher concentrations of cloud condensation nuclei (particles onto which liquid cloud droplets condense) suppress rain in shallow clouds, allowing clouds to loft more condensate through the freezing level, increasing the latent heat released when cloud water freezes, and enhancing buoyancy as soon as enough condensate precipitates out (13). The second mechanism relies on a “warm-phase” pathway: Higher aerosol concentrations reduce supersaturation in liquid clouds, increasing latent heat release through additional condensation of water vapor (14).

Here, we use idealized high-resolution simulations with the System for Atmospheric Mod-

eling (SAM) (15), scale analysis, and plume model calculations to describe a novel “humidity-entrainment” invigoration mechanism that is distinct from both the cold- and warm-phase mechanisms. Unlike the cold- and warm-phase mechanisms, which consider aerosol-induced changes in cloud processes independently from changes in the surrounding clear-air environment, the humidity-entrainment mechanism relies specifically on cloud-environment feedbacks. The two key ingredients are (i) an increase in environmental humidity in response to higher aerosol concentrations in clouds, and (ii) an increase in large-scale ascent in response to increased environmental humidity.

We represent changes in aerosol abundance in our simulations by varying a prescribed liquid cloud droplet number concentration ( $N_c$ ) in SAM’s cloud microphysics scheme (16) from  $50 \text{ cm}^{-3}$  (characteristic of pristine maritime environments) to  $800 \text{ cm}^{-3}$  (characteristic of polluted continental environments) (17). Because larger  $N_c$  inhibits rain formation in liquid clouds (18), the cold-phase invigoration mechanism could plausibly invigorate convection in simulations with higher  $N_c$ . However, we configure the microphysics scheme to allow no supersaturation in liquid clouds, thereby precluding the operation of the warm-phase invigoration mechanism.

We first consider simulations that include a parameterization of large-scale dynamics based on the weak temperature gradient (WTG) approximation (19, 20). The WTG parameterization diagnoses large-scale vertical motion that relaxes domain-average temperature profiles toward a reference profile and generates large-scale moisture convergence. We configure WTG-constrained simulations to mimic a localized aerosol anomaly embedded within a large-scale low-aerosol environment, but with otherwise identical boundary conditions, by taking the reference temperature profile from an  $N_c = 50 \text{ cm}^{-3}$  simulation of radiative-

convective equilibrium (RCE), where the environment evolves freely until reaching an equilibrium where heating by convection balances cooling by radiation.

Larger  $N_c$  increases large-scale ascent by up to  $4 \text{ cm s}^{-1}$  around 8 km (Fig. 1A). Because of constraints imposed by WTG dynamics, increasing  $N_c$  leaves environmental temperature profiles nearly unaltered (Fig. 1B). Increases in large-scale ascent with  $N_c$  are accompanied by increases in the upward mass flux in clouds (Fig. 1C), consistent with constraints imposed by steady-state energy and mass balances (20) as well as by increases in humidity and precipitation (Fig. 1D and fig. S1). High-percentile vertical velocities also increase with  $N_c$  by up to 100% in the upper troposphere (Fig. 1E). The increase in high-percentile vertical velocity is accompanied by greater areal coverage of strong convective cores, but not by greater average updraft speeds within strong convective cores (fig. S2). This suggests that increased high-percentile vertical velocities are linked to increased frequency rather than to increased speed of strong updrafts. Larger cloud mass fluxes and high-percentile vertical velocities are both signatures of convective invigoration, and these results—invigoration in simulations with a localized aerosol anomaly—are consistent with results from other modeling studies that have represented aerosol anomalies using spatially inhomogeneous RCE simulations (21, 22) rather than by coupling domain-wide aerosol changes to parameterized large-scale dynamics.

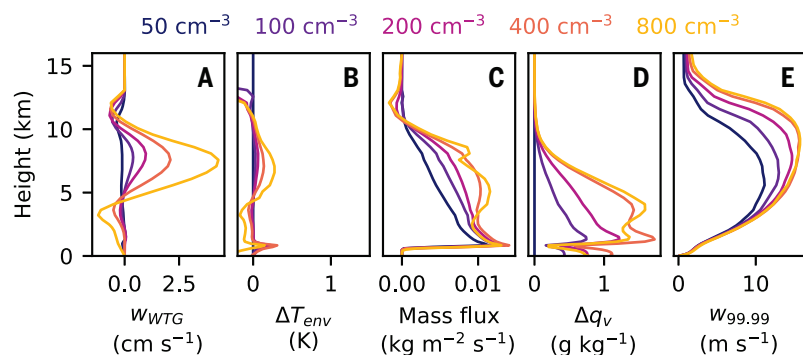
If we remove the WTG parameterization and instead run RCE simulations with varied  $N_c$  (20), the large-scale vertical velocity is constrained to vanish, and increasing  $N_c$  no longer leads to substantial changes in cloud mass fluxes or high-percentile vertical velocities (Fig. 2, C and D). Instead, similar to previous RCE studies (23), heating that would trigger large-scale ascent under WTG instead leads to increases in atmospheric temperatures (Fig. 2, A and B). Although RCE simulations lack invigoration, examining the processes that warm the atmosphere in RCE simulations provides insight into invigoration in WTG simulations. Additionally, focusing on understanding those processes in RCE simulations allows us to link changes in microphysics and atmospheric heating without the complexity introduced by parameterized large-scale circulations.

Because the cold-phase invigoration mechanism relies on the latent heat of fusion, we tested whether it is responsible for the warmer troposphere in high- $N_c$  RCE simulations by conducting mechanism-denial experiments with the latent heat of fusion  $L_f$  set to 0. Although a slightly deeper layer of the troposphere is warmed when the latent heat of fusion is non-zero, we find that differences between low- and high- $N_c$  temperature profiles with  $L_f = 0$

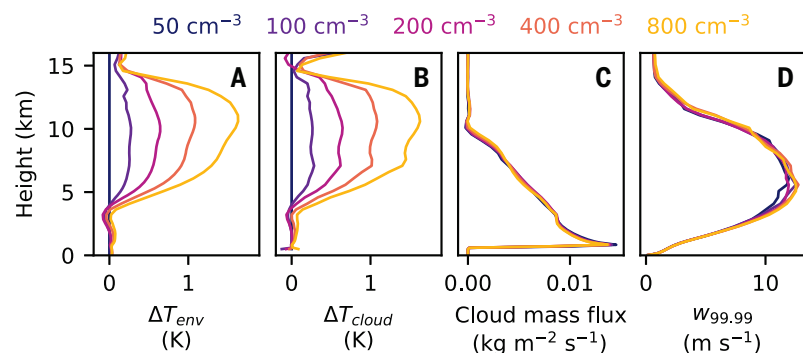
Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

\*Corresponding author. Email: thabbott@mit.edu





**Fig. 1. Invigoration in WTG simulations.** (A to E) WTG vertical velocities ( $w_{\text{WTG}}$ ) (A), domain- and time-mean temperatures ( $T_{\text{env}}$ ) (B), cloud mass fluxes (C), domain- and time-mean specific humidity ( $q_v$ ) (D), and 99.99th percentile vertical velocities ( $w_{99.99}$ ) (E) in WTG simulations with varied  $N_c$ . Domain- and time-mean temperatures and specific humidities are plotted as a difference ( $\Delta$ ) from the WTG simulation with  $N_c = 50 \text{ cm}^{-3}$ . Cloud mass fluxes include contributions from all grid points with cloud water mixing ratios above the smaller of two values:  $10^{-2} \text{ g kg}^{-1}$  or 1% of the saturation specific humidity.



**Fig. 2. Warming in RCE simulations.** (A to D) Domain- and time-mean temperatures ( $T_{\text{env}}$ ) (A), mean in-cloud temperatures ( $T_{\text{cloud}}$ ) (B), cloud mass fluxes (C), and 99.99th percentile vertical velocities ( $w_{99.99}$ ) (D) in RCE simulations with varied  $N_c$ . Temperatures are plotted as a difference ( $\Delta$ ) from the WTG simulation with  $N_c = 50 \text{ cm}^{-3}$ . In-cloud temperatures and cloud mass fluxes include contributions from all grid points with cloud water mixing ratios above the smaller of two values:  $10^{-2} \text{ g kg}^{-1}$  or 1% of the saturation specific humidity.

are almost identical to the default simulations (Fig. 3A). The persistence of the temperature differences provides strong evidence that the cold-phase mechanism is not responsible for warming the upper troposphere in high- $N_c$  simulations of RCE.

If not the latent heat of fusion, then what produces a warmer middle and upper troposphere in simulations with higher  $N_c$ ? Humidity changes are one possibility. Previous work has shown that updraft temperatures are closely linked, through entrainment, to environmental relative humidity (24). Moreover, scale analysis (20) suggests that parcel temperatures might increase by as much as 1 K per 1% change in relative humidity, whereas changes in the amount of frozen condensate are unlikely to change parcel temperatures by more than 1 K total.

A simple plume model, which links tropospheric warming to increased humidity in RCE

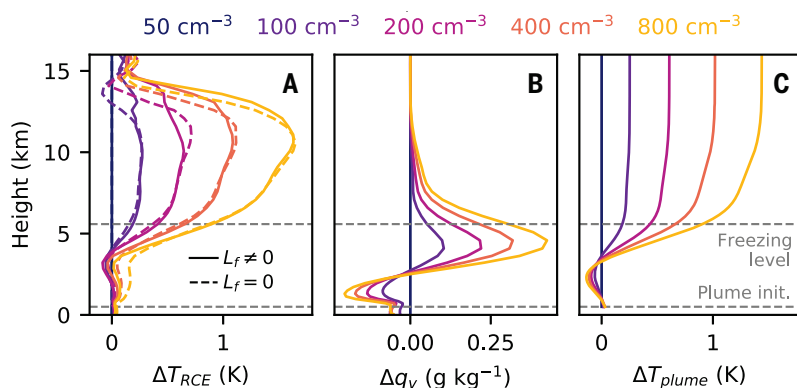
simulations, provides evidence that changes in humidity are responsible for warming the troposphere at high  $N_c$  (20). Mean specific humidity profiles in RCE simulations increase with  $N_c$  between 2.5 and 7.5 km (Fig. 3B). Because updrafts are cooled less by entrainment in a moister environment, the plume model predicts warmer temperature profiles in high- $N_c$  simulations above ~3 to 4 km, approximately coincident with the level where simulated temperature differences first become large (Fig. 3C). With an entrainment rate of  $1 \text{ km}^{-1}$ , the plume model also reproduces the magnitude of temperature differences in the upper troposphere, although this is sensitive to the choice of entrainment rate. Overall, our mechanism-denial experiments, scale analysis, and plume calculations all point to changes in atmospheric humidity as the driver of tropospheric warming at high  $N_c$ .

But why does humidity itself change in response to aerosol increases? In RCE, environmental humidity is set by a balance between moistening by air detrained from clouds and drying by clear-air subsidence (25, 26). By inhibiting the formation of rain in liquid clouds (18), increases in aerosol concentrations may increase the mass of condensate in air detrained below the freezing level, which in turn could act to increase tropospheric relative humidity by increasing detrainment moistening (26). High- $N_c$  RCE simulations show larger condensate concentrations and higher cloud evaporation rates near levels where humidity increases (Fig. 4); both features are consistent with higher humidity driven by aerosol-induced changes in detrainment moistening.

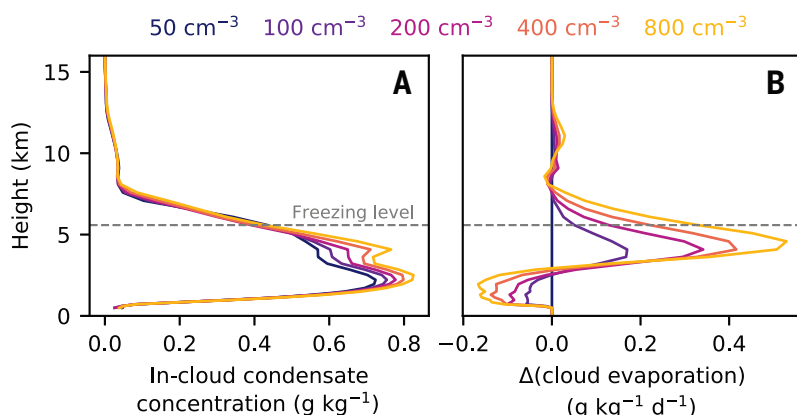
Our analysis of RCE simulations points to a causal chain for microphysical invigoration whereby greater aerosol concentrations increase atmospheric humidity, heating the troposphere as updrafts are cooled less by entrainment, and triggering large-scale ascent under WTG. Large-scale ascent imports moisture and enhances the initial humidity anomaly; differences in our simulations between humidity anomalies in RCE (Fig. 3) and WTG (fig. S1) suggest that this feedback can amplify humidity perturbations by a factor of 4 or more. The strength of this feedback is difficult to estimate a priori and may be sensitive to details of the WTG implementation. However, we expect qualitatively similar results (ascent under increased aerosol loading) as long as the initial atmospheric heating produces large-scale ascent.

Our results emphasize the importance of interactions between clouds and their environment in determining the cloud-ensemble response to changes in cloud microphysics: The “humidity-entrainment” mechanism that invigorates convection in our simulations is intrinsically linked to the role that convective clouds play in determining environmental humidity. A contemporaneous modeling study (27) also emphasizes cloud-environment feedbacks as an agent of microphysical invigoration. Like our study, theirs finds that higher  $N_c$  increases cloud evaporation, which in turn invigorates convection without the need for ice-phase microphysics. Unlike our study, however, theirs focuses on changes in convection in RCE, which responds to constraints very different from those that may affect convection in the WTG simulations described above.

In contrast to our results, a recent study (28) found only a weak link between aerosol concentrations and precipitation in a similar set of WTG-constrained simulations. The key difference is likely their focus on simulations with strong large-scale ascent forced by a sea surface temperature perturbation. Analytical models and numerical sensitivity experiments (20) (figs. S3 to S5) both suggest that humidity and



**Fig. 3. RCE mechanism-denial experiments and plume calculations.** (A) Differences ( $\Delta$ ) from an  $N_c = 50 \text{ cm}^{-3}$  control of domain- and time-mean temperature profiles from RCE simulations with default  $L_f$  (solid lines) and  $L_f = 0$  (dashed lines). (B and C) Similar plots as in (A) for domain- and time-mean specific humidity profiles from RCE simulations used as input for plume calculations (B) and environmental temperature profiles from plume calculations (C). The gray dashed line at 500 m shows the level where plume calculations are initialized with domain- and time-mean temperatures from simulations; the gray dashed line near 5 km indicates the freezing level, calculated as the lowest model level where more than 5% of the mean cloud water mass is ice.



**Fig. 4. In-cloud condensate concentration and cloud evaporation in RCE simulations.** (A) The in-cloud condensate concentration is calculated as a conditional average over all grid cells where the liquid plus ice cloud mass concentration is larger than the smaller of two values:  $10^{-2} \text{ g kg}^{-1}$  and 1% of the saturation specific humidity. (B) The cloud evaporation rate is calculated as an average over all grid cells and plotted as a difference ( $\Delta$ ) from the  $N_c = 50 \text{ cm}^{-3}$  simulation. The gray dashed line indicates the freezing level calculated as in Fig. 3.

precipitation become less sensitive to changes in detrainment (and thus to changes in aerosol concentration) in the presence of strong large-scale ascent.

In the tropics, updraft speeds appear highest in regions with a moderately dry troposphere and not in regions where large-scale ascent pushes the atmosphere toward saturation (7). Our numerical sensitivity experiments are consistent with this observation: Average updraft speeds in strong convective cores are highest in simulations with weak ascent or moderate subsidence (fig. S6). Because relative humidity is most sensitive to changes in detrainment when ascent is weak,

the humidity-entrainment mechanism is likely to increase the frequency of strong updrafts most effectively in regions with a moderately dry troposphere and fast updrafts, although humidity increases may also lead to an accompanying decrease in peak updraft speeds.

Finally, because the humidity-entrainment mechanism relies on a three-way link between aerosols, humidity, and convective vigor—all of which are either directly or indirectly observable—our results provide a target for future observational work. Although the connections found in our idealized models may be challenging to detect in observations, doing so would point to a key pathway by

which aerosols (including those produced by human activity) modify weather in the tropics.

## REFERENCES AND NOTES

1. M. O. Andreae *et al.*, *Science* **303**, 1337–1342 (2004).
2. I. Koren, Y. J. Kaufman, D. Rosenfeld, L. A. Remer, Y. Rudich, *Geophys. Res. Lett.* **32**, L14828 (2005).
3. I. Koren, L. A. Remer, O. Altartatz, J. V. Martins, A. Davidi, *Atmos. Chem. Phys.* **10**, 5001–5010 (2010).
4. Z. Li *et al.*, *Nat. Geosci.* **4**, 888–894 (2011).
5. F. Niu, Z. Li, *Atmos. Chem. Phys.* **12**, 8491–8498 (2012).
6. E. Williams, S. Stanfill, *C. R. Phys.* **3**, 1277–1292 (2002).
7. E. J. Zipser, D. J. Cecil, C. Liu, S. W. Nesbitt, D. P. Yorty, *Bull. Am. Meteorol. Soc.* **87**, 1057–1072 (2006).
8. J. A. Thornton, K. S. Virts, R. H. Holzworth, T. P. Mitchell, *Geophys. Res. Lett.* **44**, 9102–9111 (2017).
9. O. Altartatz, I. Koren, L. A. Remer, E. Hirsch, *Atmos. Res.* **140–141**, 38–60 (2014).
10. D. Rosenfeld, S. Sherwood, R. Wood, L. Donner, *Science* **343**, 379–380 (2014).
11. W. K. Tao, J. P. Chen, Z. Li, C. Wang, C. Zhang, *Rev. Geophys.* **50**, RG2001 (2012).
12. J. Fan, Y. Wang, D. Rosenfeld, X. Liu, *J. Atmos. Sci.* **73**, 4221–4252 (2016).
13. D. Rosenfeld *et al.*, *Science* **321**, 1309–1313 (2008).
14. J. Fan *et al.*, *Science* **359**, 411–418 (2018).
15. M. F. Khairoutdinov, D. A. Randall, *J. Atmos. Sci.* **60**, 607–625 (2003).
16. H. Morrison, J. A. Curry, V. I. Khvorostyanov, *J. Atmos. Sci.* **62**, 1665–1677 (2005).
17. D. Barahona, R. Sotiropoulou, A. Nenes, *J. Geophys. Res.* **116**, D09203 (2011).
18. B. A. Albrecht, *Science* **245**, 1227–1230 (1989).
19. A. H. Sobel, J. Nilsson, L. M. Polvani, *J. Atmos. Sci.* **58**, 3650–3665 (2001).
20. See supplementary materials.
21. H. Morrison, W. W. Grabowski, *J. Atmos. Sci.* **70**, 3533–3555 (2013).
22. P. N. Blossey, C. S. Bretherton, J. A. Thornton, K. S. Virts, *Geophys. Res. Lett.* **45**, 9305–9313 (2018).
23. J. T. Seeley, D. M. Roms, *Geophys. Res. Lett.* **43**, 3572–3579 (2016).
24. M. S. Singh, P. A. O’Gorman, *Geophys. Res. Lett.* **40**, 4398–4403 (2013).
25. D. M. Roms, *J. Clim.* **27**, 7432–7449 (2014).
26. M. S. Singh, R. A. Warren, C. Jakob, *J. Adv. Model. Earth Syst.* **11**, 3973–3994 (2019).
27. X. R. Chua, Y. Ming, ESSOAr 10502188.2 [preprint] (9 November 2020).
28. U. M. Anber, S. Wang, P. Gentine, M. P. Jensen, *J. Atmos. Sci.* **76**, 2885–2897 (2019).
29. T. H. Abbott, T. W. Cronin, Data for “Aerosol invigoration of atmospheric convection through increases in humidity” (version 2), Zenodo (2020). <https://doi.org/10.5281/zenodo.4071888>.

## ACKNOWLEDGMENTS

We thank X. R. Chua, Y. Ming, and R. Rousseau-Rizzi for discussions about this work; J. Wilcots and H. Drake for comments on the manuscript; M. Khairoutdinov for providing SAM; and P. Blossey and an anonymous reviewer for constructive reviews. **Funding:** T.H.A. and T.W.C. were supported by NSF AGS 1740533: Convection and Rainfall Enhancement over Mountainous Tropical Islands. **Author contributions:** T.H.A. and T.W.C. designed research and wrote the paper; T.H.A. performed research and analyzed data. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** SAM is publicly available (<http://rossby.msfc.suunysb.edu/~marat/SAM.html>). Modifications to the SAM source code, simulation input files, and simulation output data underlying this work are archived at (29).

## SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/371/6524/83/suppl/DC1  
Materials and Methods  
Supplementary Text  
Figs. S1 to S8  
References (30–38)

28 April 2020; accepted 24 November 2020  
10.1126/science.abc5181

## PROTEIN FOLDING

## Evolution of fold switching in a metamorphic protein

Acacia F. Dishman<sup>1,2</sup>, Robert C. Tyler<sup>1</sup>, Jamie C. Fox<sup>1</sup>, Andrew B. Kleist<sup>1,2</sup>, Kenneth E. Prehoda<sup>3</sup>, M. Madan Babu<sup>4,5</sup>, Francis C. Peterson<sup>1</sup>, Brian F. Volkman<sup>1\*</sup>

Metamorphic proteins switch between different folds, defying the protein folding paradigm. It is unclear how fold switching arises during evolution. With ancestral reconstruction and nuclear magnetic resonance, we studied the evolution of the metamorphic human protein XCL1, which has two distinct folds with different functions, making it an unusual member of the chemokine family, whose members generally adopt one conserved fold. XCL1 evolved from an ancestor with the chemokine fold. Evolution of a dimer interface, changes in structural constraints and molecular strain, and alteration of intramolecular protein contacts drove the evolution of metamorphosis. Then, XCL1 likely evolved to preferentially populate the noncanonical fold before reaching its modern-day near-equal population of folds. These discoveries illuminate how one sequence has evolved to encode multiple structures, revealing principles for protein design and engineering.

**M**etamorphic proteins defy the protein folding paradigm, in which each amino acid sequence adopts one defined fold (monomorphic). They switch reversibly between entirely distinct folds, often with different functions, on the time scale of seconds (1). Metamorphic proteins undergo large-scale structural changes (e.g., alterations in secondary structure and hydrogen bonding networks), whereas allosteric proteins exhibit smaller-scale conformational dynamics (1, 2). Whereas only about six metamorphic proteins have been well characterized (2–4), estimates suggest that metamorphic proteins may constitute up to 4% of proteins in the Protein Data Bank (PDB) (5). The emergence of new protein folds on evolutionary time scales has been examined (6–9), but it remains unclear how metamorphic folding evolves in a single protein.

Among metamorphic proteins, the human chemokine XCL1 (lymphotactin) undergoes one of the most pronounced structural switches, involving complete rearrangement of hydrogen bonding networks (Fig. 1, A and B) (10). XCL1 belongs to a family of 46 human chemokines: small, secreted proteins that direct immune cell migration. Whereas non-XCL1 chemokines adopt a monomorphic, conserved  $\alpha\beta$  fold (chemokine fold) (11), XCL1 reversibly interconverts between the chemokine fold and a dimeric, all- $\beta$  fold (alternate fold) with no structural similarity to other known proteins (Fig. 1A) (10). Non-XCL1 chemokines execute two essential functions using one structure, but XCL1 divides these two functions between its two folds: The chemokine structure activates XCL1's cognate G protein-coupled recep-

tor (GPCR), whereas the alternate structure binds glycosaminoglycans (GAGs) (10). Additionally, the alternate structure is directly antimicrobial, like a subset of chemokines (12). In a family of proteins that adopt a single fold, how did XCL1 evolve to become metamorphic?

To better understand the evolution of metamorphosis in XCL1, we inferred its phylogenetic tree by using ancestral reconstruction (Fig. 1D and fig. S1). This technique utilizes multiple sequence alignments of modern-day proteins from different species to infer amino acid sequences of shared ancestors (13) and has been used to investigate various evolutionary questions (14–19). We created a multiple alignment for 457 chemokine sequences from 30 vertebrate species incorporating 14 chemokine structures (20), from which we inferred the phylogeny of XCL1 and related chemokines using maximum likelihood methods [i.e., the inferred ancestral sequences have the highest probability of producing the modern-day sequences (13, 21, 22)]. Phylogenetic analysis indicates that XCL1 is most closely related to another (monomorphic) chemokine, CCL20 (Fig. 1D and fig. S1) (12). We resurrected (i.e., expressed and purified) the last shared ancestor of XCL1 and CCL20 (Anc.0). Unlike the temperature- and salt-dependent structural equilibrium of XCL1 (23), the Anc.0 heteronuclear single-quantum coherence (HSQC) spectrum is unchanged from 10° to 50°C, with or without NaCl (fig. S2). We solved the Anc.0 nuclear magnetic resonance (NMR) structure: It adopts the canonical chemokine fold (Fig. 1C and table S1). Despite sharing only 40% sequence identity with extant human XCL1, Anc.0's structure is highly similar to human XCL1's chemokine fold (root mean square deviation of 1.45 Å) (Fig. 1, A and C). Together, these data suggest that a metamorphic protein evolved from a monomorphic ancestor.

Non-XCL1 chemokines have two conserved disulfide bonds, one of which is incompatible with the alternate fold of XCL1. Restoring

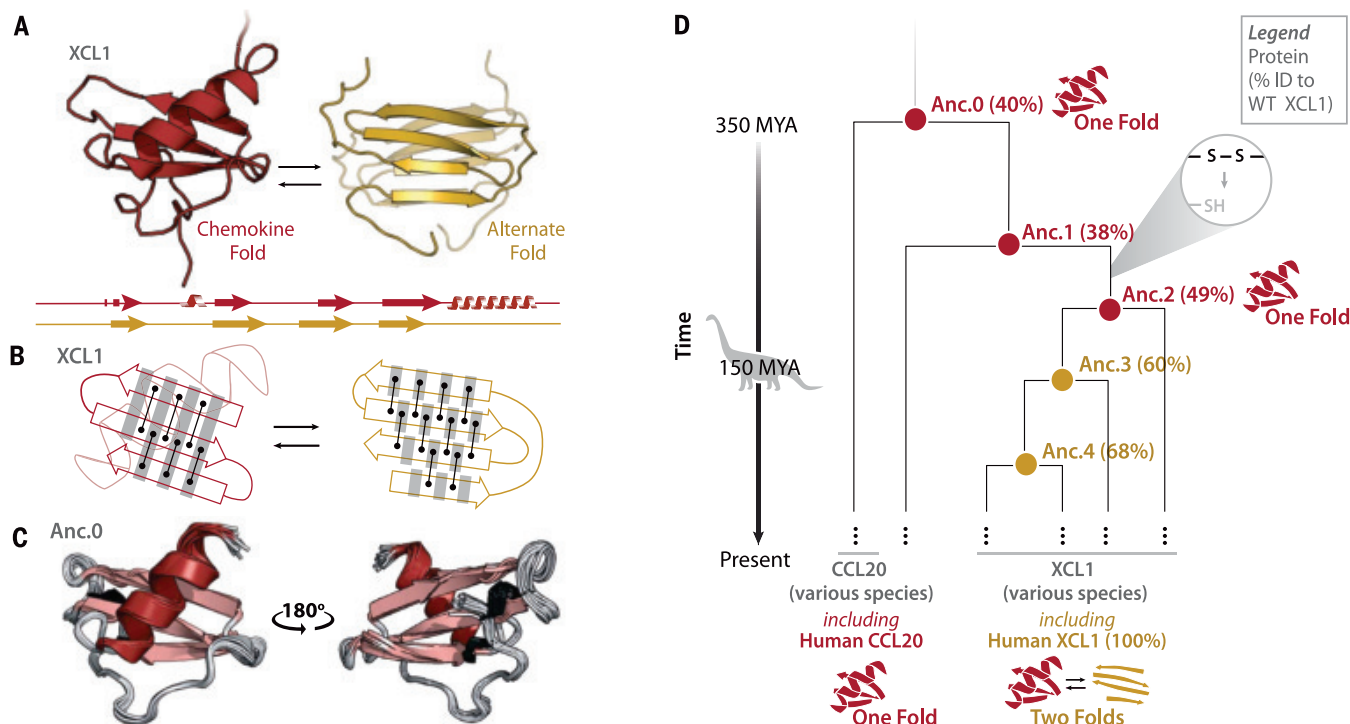
the missing disulfide bond to XCL1 makes it monomorphic, locked into the chemokine fold (24). Anc.0 has both chemokine disulfide bonds, so we sought to identify the interval where one disulfide was likely lost, because this may have imparted metamorphism. Maximum likelihood methods (13, 21, 22) identified several nodes along the evolutionary trajectory from Anc.0 to extant XCL1. By locating the oldest node whose sequence lacks the cysteines needed for one of the disulfides, we identified the interval where one disulfide was lost (Fig. 1D), culminating in an ancestor called Anc.2 (Fig. 1D). Anc.2 adopts a single fold at all temperature and salt conditions tested by NMR (10° to 50°C, with or without NaCl), indicating that disulfide loss alone did not enable metamorphism (Fig. 1D and fig. S3).

Given that loss of a conserved disulfide was necessary but insufficient to impart metamorphism to XCL1, what evolutionary changes resulted in the emergence of metamorphosis? We resurrected the remaining inferred ancestral sequences (Anc.3 and Anc.4) and found by NMR that they both interconvert reversibly between two distinct folds (Fig. 2, A to C, and figs. S4 to S7), despite sharing only 60 and 68% sequence identity with human XCL1, respectively. This suggests that a wide range of sequences can encode metamorphism, because here metamorphism is encoded by sequences that differ at up to 40% of positions. ZZ-exchange experiments reveal that Anc.3 [rate of exchange ( $k_{ex}$ ), 0.97 s<sup>-1</sup>] and Anc.4 ( $k_{ex}$ , 2.6 s<sup>-1</sup>) exchange on the time scale of seconds, similar to human XCL1 [ $k_{ex}$ , 0.90 s<sup>-1</sup> (25)], consistent with XCL1 evolving to remain metamorphic over hundreds of millions of years, balancing occupancy of its two structures by keeping exchange rates in a narrow range.

Because extant human XCL1 occupies its two folds in equal proportion, we sought to determine whether Anc.3 and Anc.4 do also. HSQC peak intensities can quantify the fractional population of the different folds of XCL1, which is determined by the concentration-independent equilibrium constant for fold switching. HSQC experiments show that under near-physiologic conditions (37°C, 150 mM NaCl) at identical protein concentrations, 92 ± 3.7% of the Anc.3 population occupies the chemokine fold (Fig. 2, A, B, and D, and table S2). However, only 9.3 ± 2.0% of the Anc.4 population occupies the chemokine fold (Fig. 2, A, B, and D, and table S2). To confirm the robustness of these results to statistical uncertainty in ancestrally reconstructed sequences, we expressed and purified “Alt.” ancestral proteins, where we replaced every residue with the next most likely residue if the next most likely residue was predicted with >20% probability, creating a “worst-case” ancestor with the alternate amino acid at all positions, as is the standard in the field (26) (fig. S8). Consistent

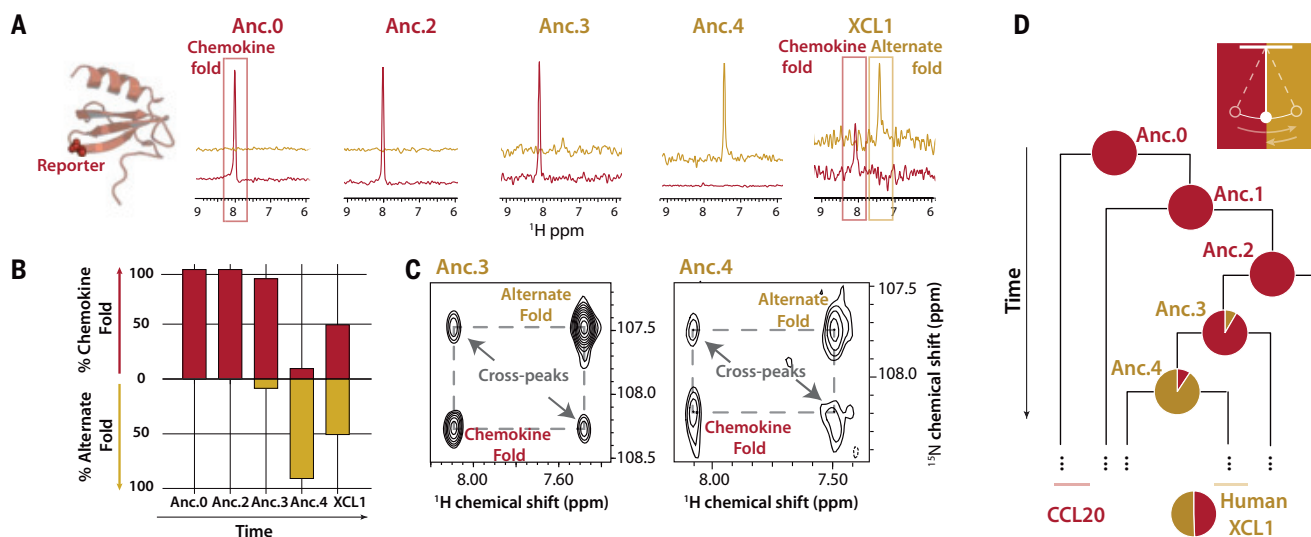
<sup>1</sup>Department of Biochemistry, Medical College of Wisconsin, Milwaukee, WI, USA. <sup>2</sup>Medical Scientist Training Program, Medical College of Wisconsin, Milwaukee, WI, USA. <sup>3</sup>Institute of Molecular Biology, Department of Chemistry and Biochemistry, University of Oregon, Eugene, OR, USA. <sup>4</sup>MRC Laboratory of Molecular Biology, Cambridge, UK. <sup>5</sup>Department of Structural Biology and Center for Data Driven Discovery, St. Jude Children's Research Hospital, Memphis, TN, USA.  
\*Corresponding author. Email: bvolkman@mcw.edu





**Fig. 1. Evolutionary history of XCL1.** (A) Structures and secondary structure diagrams for XCL1's chemokine fold (red) and alternate fold (gold). (B) Cartoon comparing hydrogen bonding networks in each fold's  $\beta$  strands. Black lines between dots represent pairs of hydrogen bonds between amino acids. Gray shading highlights the same set of residues. Interconversion between XCL1's chemokine fold and alternate fold requires the  $\beta 2$  strand to rotate  $180^\circ$  and shift by one residue relative to the  $\beta 1$  and  $\beta 3$  strands, establishing an entirely new

hydrogen bonding pattern. (C) Ensemble of the top 20 NMR structures for Anc.0 (PDB ID 7JH1), colored by secondary structure. Disulfide bonds are shown as dark gray sticks. (D) Simplified phylogenetic tree showing XCL1's evolutionary history, beginning with the last shared ancestor (Anc.0) of XCL1 and another chemokine (CCL20). MYA, million years ago. Nodes represent reconstructed ancestral sequences. For each ancestral sequence, the percent identity to extant human XCL1 is shown in parentheses.



**Fig. 2. Evolutionary progression of metamorphic folding in XCL1.** (A) One-dimensional traces (from HSQC experiments performed at  $37^\circ\text{C}$ , 150 mM NaCl, 350  $\mu\text{M}$  protein concentration) for a glycine peak (Gly<sup>44</sup> in XCL1) that is diagnostic of XCL1's metamorphic conformation, shown as spheres on Anc.0's structure at left. Because Gly<sup>44</sup> occupies distinct chemical environments in XCL1's different native structures, there are two Gly<sup>44</sup> peaks with chemical shifts of  $\sim 8.0$   $^1\text{H}$  ppm (chemokine fold, red) and  $\sim 7.4$   $^1\text{H}$  ppm (alternate fold, gold). (B) Fractional abundances of the chemokine and alternate folds were calculated as averages of relative HSQC peak volumes for two reporter

residues (Gly<sup>44</sup> and the indole NH of Trp<sup>55</sup> in XCL1). All spectra were collected at  $37^\circ\text{C}$  with 150 mM NaCl and 350  $\mu\text{M}$  protein concentration. (C) ZZ-exchange dynamic analysis (46) of Anc.3 at  $40^\circ\text{C}$  and Anc.4 at  $25^\circ\text{C}$ , both with 0 mM NaCl, using the glycine reporter. Cross peaks indicate the presence of structural interconversion. Conditions for each ancestor were chosen to maximize the presence of both folds to enhance detection of cross peaks. (D) Simplified phylogenetic tree indicating the back-and-forth evolutionary trajectory of XCL1 metamorphosis. Pie charts present fractional occupancies of the chemokine fold (red) and the alternate fold (gold).

with our results for the maximum likelihood ancestors, Alt.Anc.0 and Alt.Anc.2 are monomeric, whereas Alt.Anc.3 and Alt.Anc.4 are metamorphic (fig. S8), and Alt.Anc.4 populates the chemokine fold at  $11 \pm 1.4\%$ . This suggests that XCL1 likely evolved from a single-fold ancestor (Anc.2) to a metamorphic ancestor that prefers the chemokine fold (Anc.3), then to a metamorphic ancestor that prefers the alternate fold (Anc.4), and finally to an extant metamorph that equally populates two different folds (Fig. 2, B and D).

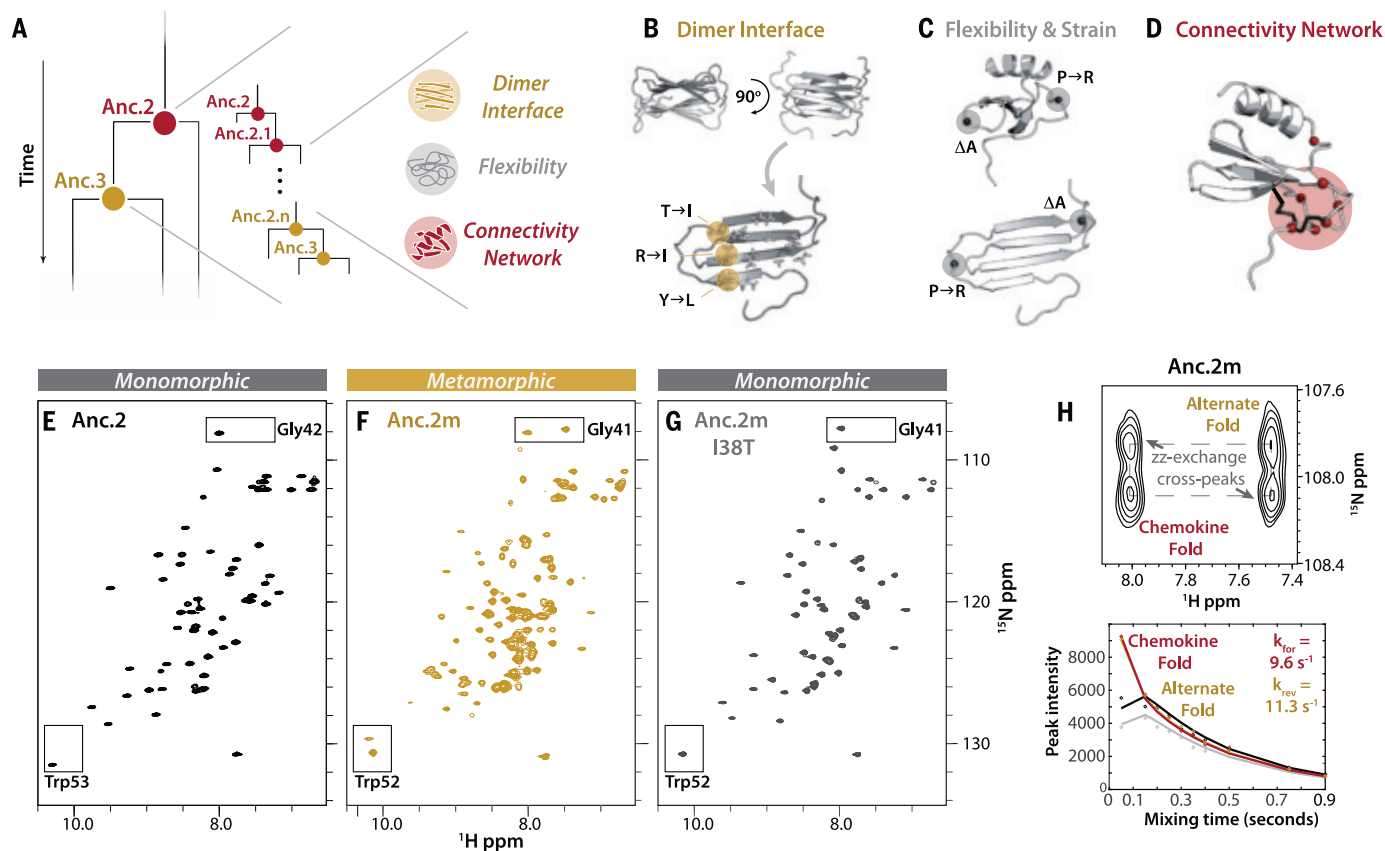
The aligned sequences of Anc.2 and Anc.3 differ in 26 positions distributed throughout sequence and structure (fig. S1C). To uncover how these sequence changes introduced metamorphosis, we examined these positions on the basis of three criteria: sequence comparison of positions at the alternate structure's dimer interface, sequence comparison of positions likely to affect overall structural flexibility, and analysis of residue-residue contacts in the chemokine structure (Fig. 3A). Another

sequence-based criterion that has been used to predict protein metamorphosis is inaccurate secondary structure prediction (5, 27); however, each ancestral sequence and wild-type (WT) XCL1 are predicted to adopt the secondary structure profile of the chemokine fold (fig. S9).

Several metamorphic proteins form protein-protein interfaces likely to stabilize one of their two structures, including RfaH (28), Selecase (29), and XCL1 (23). For XCL1, the alternate structure forms a stable dimer interface that is likely important for metamorphic folding, so we performed sequence comparison of Anc.2 and Anc.3 at XCL1 dimer interface positions. We found that three interface positions switch from polar or charged in Anc.2 to aliphatic side chains in Anc.3 [Tyr<sup>11</sup>→Ile (Y11I), T38I, and R43I] (Fig. 3B), making contacts with the rest of the apolar dimer interface more favorable. Sequence comparison also identified two changes between Anc.2 and Anc.3 that likely increase structural flexibility

(P15R) or create strain (a deletion in the  $\beta$ 1- $\beta$ 2 loop at position 29) in the chemokine fold (Fig. 3C).

Networks of noncovalent residue-residue contacts are key indicators of protein conformation and stability (30, 31) that mediate essential biological phenomena (32–34). Because metamorphosis likely relies on a nuanced balance of conformational stability across multiple protein folds, we suspected that changes in noncovalent contact networks might contribute to the evolution of protein metamorphosis. We compared contact networks in human XCL1 versus Anc.0 and identified seven positions that make more than three contacts in Anc.0 and fewer than three contacts in XCL1 and whose identities differ between Anc.2 and Anc.3. We focused on these sequence substitutions because they decreased connectivity in Anc.0 (potentially permissive for interconversion) but minimized changes in connectivity in XCL1 (preserving the chemokine fold) (30) (Fig. 3D and fig. S10), suggesting that mutations



**Fig. 3. Metamorphic folding is enabled by the combined presence of changes in dimer interface, structural flexibility, and residue contact networks.** (A) Schematic of the hypothesis that metamorphosis evolved over time via a subset of key sequence changes from Anc.2 toward Anc.3. (B) Dimer interface positions that are polar in Anc.2 and hydrophobic in Anc.3. (C) Positions likely to be important for structural flexibility and creation of strain in the chemokine fold. Amino acid abbreviations in (B) and (C): I, isoleucine; L, leucine; P, proline; R, arginine; T, threonine; Y, tyrosine. (D) Positions selected

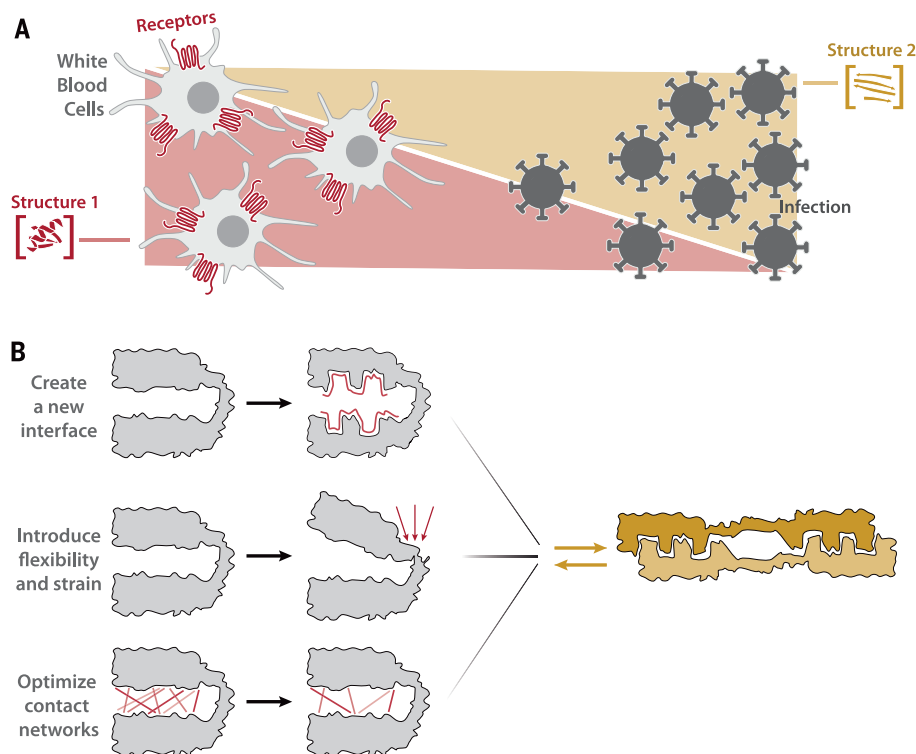
for replacement in Anc.2 with contact network analysis (C $\alpha$ , red spheres). Anc.0 disulfides are shown in black. (E to G) HSQC spectra for Anc.2 (E), Anc.2m (F), and Anc.2m I38T (G) (50°C, 20 mM NaPO<sub>4</sub>, pH 6.0). Boxes indicate conformational reporter residues. (H) ZZ-exchange analysis for Anc.2m (50°C, 20 mM NaPO<sub>4</sub>, pH 6.0) with reporter residue Gly<sup>41</sup>. (Top) Cross peaks indicate interconversion. (Bottom) Curves fit to peak intensities versus time to calculate kinetic parameters.  $k_{\text{forward}}$  ( $k_{\text{for}}$ ), interconversion from the alternate fold toward the chemokine fold;  $k_{\text{reverse}}$  ( $k_{\text{rev}}$ ), interconversion in the reverse direction.

that permit rewiring of these contacts could enable metamorphic folding. These positions largely cluster around Anc.0's second disulfide bond (Fig. 3D and fig. S10C), "gluing" its structure in place. After the disulfide was lost, this remaining structural glue "dissolved" as metamorphosis evolved. Position 27, which points into the glue region, is a glutamine in Anc.0, Anc.2, and Anc.3, but in Anc.4 it is a bulky lysine, which is likely to sterically destabilize the chemokine fold, perhaps partly accounting for the fact that Anc.4 prefers the alternate fold. In human XCL1, this position is a threonine, perhaps facilitating XCL1's shift to a 50/50 structural equilibrium.

To identify sequence changes sufficient to introduce metamorphic folding, we mutated positions in Anc.2 (monomorphic) to match Anc.3 (metamorphic) on the basis of the findings described above (table S3). We evaluated the resulting variants for fold switching via HSQC experiments. We found that combining the three dimer interface substitutions, two flexibility-altering substitutions, and seven connectivity-altering substitutions (11 of 26 possible sequence changes; Anc.2m) (Fig. 3, A to D, and table S3) conferred the ability to populate the alternate fold (fig. S11 and Fig. 3, E to H). Anc.2 variants incorporating all three dimer interface substitutions alone, the two flexibility-altering changes alone, or the seven connectivity-altering changes alone are not metamorphic and only occupy the chemokine fold (table S3), suggesting that these three sets of changes had to occur in combination to introduce metamorphic folding. Reverting the least disruptive mutation at the dimer interface (I38T) in Anc.2m disables metamorphosis, causing Anc.2m I38T to adopt exclusively the chemokine fold, indicating that the T38I mutation is necessary but insufficient for metamorphic folding (Fig. 3G and table S3). This highlights how mutations that stabilize an alternate fold through formation of a protein-protein interface can critically influence metamorphic folding, as is seen in other metamorphic proteins (28). Together, these findings show that concurrent changes in interaction interfaces, structural flexibility, and noncovalent residue-residue contact networks act together to drive the evolution of protein metamorphosis.

Lattman and Rose proposed that a protein's specific folded structure can be encoded either by amino acids distributed throughout its sequence (distributed control) or discrete sites (centralized control) (35). Whereas both distributed (36) and centralized (37) control of folding have been described, the requirement for numerous, structurally distant mutations of different kinds to enable metamorphosis suggests that distributed control governs XCL1's structure.

Despite growing interest in metamorphic proteins (2–5), the mechanism by which one



**Fig. 4. Functional advantages and design insights provided by protein metamorphosis.** (A) Metamorphosis enables spatio-temporal control of protein function through modulation of the relative population of two different folded states in a concentration-dependent manner. For example, XCL1 can populate the antimicrobial fold close to a site of infection and the GPCR binding fold further away. (B) The emergence in nature or de novo design of fold-switching proteins may require creation of a protein-protein interface that stabilizes the alternate fold (top), incorporation of molecular flexibility and strain (middle), and optimization of residue-residue contact networks to avoid thermodynamic or kinetic trapping in a single fold (bottom).

sequence encodes two structures remains unclear. Here, ancestral reconstruction illuminates the molecular evolution of protein metamorphosis in XCL1. Metamorphic proteins have been hypothesized to represent evolutionary bridges, or "snapshots" of proteins "caught in the act" of evolving a new fold (8, 38, 39). Were this the case for XCL1, its ancestors would likely have progressed from populating only the chemokine fold to adopting exclusively the alternate fold. Instead, our data suggest that human XCL1 likely evolved from a metamorphic ancestor that may preferentially adopt the chemokine fold to one that may preferentially adopt the alternate fold, before evolving to occupy both structures in approximately equal proportion. This suggests that XCL1 is not evolving from one fold to a new one but is evolving to remain metamorphic, indicating that metamorphosis may be a molecular phenotype that was selected for in XCL1 rather than a transient feature of an evolutionary intermediate.

Why would metamorphosis be favored? All chemokines activate GPCRs and bind GAGs, and some chemokines, including XCL1, are directly antimicrobial (12, 40–43). Non-XCL1 chemokines carry out these three functions

using one fold, whereas XCL1's chemokine fold activates its cognate GPCR and its alternate fold binds GAGs and is directly antimicrobial (10, 12, 40). As such, metamorphic folding could confer dynamic control over the fractional population, and therefore the activity, of each structure (Fig. 4A), avoiding the need to transcribe and translate multiple genes and degrade or inhibit multiple proteins to turn multiple functions on and off, thus allowing for fast switching between functional states mediated by distinct folds. It could also enhance a specific function (e.g., GAG binding) or permit acquisition of a new function (e.g., antimicrobial activity). Specifically, metamorphic folding may enable XCL1 to kill pathogens directly at the site of infection (alternate fold) and stimulate antigen cross-presentation to effector cells (chemokine fold), coordinating humoral and cell-mediated immune responses and allowing for spatiotemporal regulation of multiple functions (Fig. 4A). If fold switching thus enhances the ability of other metamorphic proteins to carry out their biologic roles, then metamorphic proteins may be more common than previously thought. Why metamorphic folding evolved in XCL1 and not other chemokines, however, remains mysterious.



Analysis of proteins like XCL1 that reversibly switch between two well-defined folds can inform the de novo design of fold-switching proteins, which remains a challenge. Our work suggests that designed fold-switching proteins may need to incorporate protein-protein interfaces that stabilize one fold, structural constraints and strain to enable interconversion, and residue-residue contact networks that allow stable formation of both folds without trapping the protein in either fold (Fig. 4B). LOCKR, a designed protein switch system currently being developed for therapeutic use, has already been constructed by stabilizing a protein-protein interface, demonstrating the utility of this approach (44, 45). Moreover, the broad span of XCL1 sequence space compatible with metamorphic interconversion implies that design of fold-switching proteins could be within reach, especially with the help of principles uncovered in this study.

## REFERENCES AND NOTES

1. A. G. Murzin, *Science* **320**, 1725–1726 (2008).
2. A. F. Dishman, B. F. Volkman, *ACS Chem. Biol.* **13**, 1438–1446 (2018).
3. M. Lella, R. Mahalakshmi, *Biochemistry* **56**, 2971–2984 (2017).
4. H. Zamora-Carreras, B. Maestro, J. M. Sanz, M. A. Jiménez, *ChemBioChem* **21**, 432–441 (2020).
5. L. L. Porter, L. L. Looger, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 5968–5973 (2018).
6. S. G. Peisajovich, L. Rockah, D. S. Tawfik, *Nat. Genet.* **38**, 168–174 (2006).
7. N. Tokuriki, D. S. Tawfik, *Curr. Opin. Struct. Biol.* **19**, 596–604 (2009).
8. I. Yadid, N. Kirshenbaum, M. Sharon, O. Dym, D. S. Tawfik, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 7287–7292 (2010).
9. N. Tokuriki, D. S. Tawfik, *Science* **324**, 203–207 (2009).
10. R. L. Tuinstra et al., *Proc. Natl. Acad. Sci. U.S.A.* **105**, 5057–5062 (2008).
11. E. S. Kuloğlu et al., *Biochemistry* **40**, 12486–12496 (2001).
12. A. M. Nevins et al., *Biochemistry* **55**, 3784–3793 (2016).
13. Z. Yang, S. Kumar, M. Nei, *Genetics* **141**, 1641–1650 (1995).
14. B. S. Chang, M. A. Kazmi, T. P. Sakmar, *Methods Enzymol.* **343**, 274–294 (2002).
15. J. W. Thornton, E. Need, D. Crews, *Science* **301**, 1714–1717 (2003).
16. G. C. Finnigan, V. Hanson-Smith, T. H. Stevens, J. W. Thornton, *Nature* **481**, 360–364 (2012).
17. B. S. Chang, K. Jönsson, M. A. Kazmi, M. J. Donoghue, T. P. Sakmar, *Mol. Biol. Evol.* **19**, 1483–1489 (2002).
18. D. P. Anderson et al., *eLife* **5**, e10147 (2016).
19. D. S. Whitney, B. F. Volkman, K. E. Prehoda, *J. Am. Chem. Soc.* **138**, 15150–15156 (2016).
20. J. Pei, B. H. Kim, N. V. Grishin, *Nucleic Acids Res.* **36**, 2295–2300 (2008).
21. S. Guindon, F. Delsuc, J. F. Dufayard, O. Gascuel, *Methods Mol. Biol.* **537**, 113–137 (2009).
22. S. Guindon, O. Gascuel, *Syst. Biol.* **52**, 696–704 (2003).
23. E. S. Kuloğlu, D. R. McCaslin, J. L. Markley, B. F. Volkman, *J. Biol. Chem.* **277**, 17863–17870 (2002).
24. R. L. Tuinstra, F. C. Peterson, E. S. Elgin, A. J. Pelzek, B. F. Volkman, *Biochemistry* **46**, 2564–2573 (2007).
25. R. C. Tyler, J. C. Wieting, F. C. Peterson, B. F. Volkman, *Biochemistry* **51**, 9067–9075 (2012).
26. G. M. Eick, J. T. Bridgham, D. P. Anderson, M. J. Harms, J. W. Thornton, *Mol. Biol. Evol.* **34**, 247–261 (2017).
27. S. Mishra, L. L. Looger, L. L. Porter, *Protein Sci.* **28**, 1487–1493 (2019).
28. B. M. Burmann et al., *Cell* **150**, 291–303 (2012).
29. M. López-Pelegrín et al., *Angew. Chem.* **126**, 10800–10806 (2014).
30. M. Kayikci et al., *Nat. Struct. Mol. Biol.* **25**, 185–194 (2018).
31. L. H. Greene, *Brief. Funct. Genomics* **11**, 469–478 (2012).
32. A. Sente et al., *Nat. Struct. Mol. Biol.* **25**, 538–545 (2018).
33. T. Flock et al., *Nature* **545**, 317–322 (2017).
34. T. Flock et al., *Nature* **524**, 173–179 (2015).

35. E. E. Lattman, G. D. Rose, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 439–441 (1993).
36. K. V. Eaton et al., *Protein Eng. Des. Sel.* **28**, 241–250 (2015).
37. K. L. Stewart, E. D. Dodds, V. H. Wysocki, M. H. Cordes, *Protein Sci.* **22**, 641–649 (2013).
38. V. K. Kumirov et al., *Protein Sci.* **27**, 1767–1779 (2018).
39. M. H. J. Cordes, R. E. Burton, N. P. Walsh, C. J. McKnight, R. T. Sauer, *Nat. Struct. Biol.* **7**, 1129–1132 (2000).
40. A. F. Dishman et al., *ACS Infect. Dis.* **6**, 1204–1213 (2020).
41. L. T. Nguyen, H. J. Vogel, *Front. Immunol.* **3**, 384 (2012).
42. K. Hieshima et al., *J. Immunol.* **170**, 1452–1461 (2003).
43. D. Yang et al., *J. Leukoc. Biol.* **74**, 448–455 (2003).
44. R. A. Langan et al., *Nature* **572**, 205–210 (2019).
45. A. H. Ng et al., *Nature* **572**, 265–269 (2019).
46. N. A. Farrow, O. Zhang, J. D. Forman-Kay, L. E. Kay, *J. Biomol. NMR* **4**, 727–734 (1994).

## ACKNOWLEDGMENTS

We thank C. Peterson for helpful suggestions regarding the figures and text. **Funding:** This work was supported in part by a grant from the State of Wisconsin Tax Check-Off Program for Cancer Research and the Medical College of Wisconsin Cancer Center and NIH grants R56 AI103225, R37 AI058072, and S10 OD020000 (to B.F.V.), F30 CA236182 (to A.F.D.), and F30 CA196040 (to A.B.K.). A.F.D. and A.B.K. are members of the NIH-supported (T32 GM080202) Medical Scientist Training Program at the Medical College of Wisconsin. M.M.B. acknowledges the MRC (MC\_U105185859) and ALSAC for funding. **Author contributions:** A.F.D., R.C.T., J.C.F., F.C.P.,

K.E.P., M.M.B., and B.F.V. conceived and planned experiments. R.C.T. and K.E.P. performed ancestral sequence reconstruction. R.C.T. and F.C.P. solved the structure of Anc.O. A.F.D., R.C.T., and J.C.F. produced and purified proteins. A.F.D., R.C.T., J.C.F., and F.C.P. performed HSQC and ZZ-exchange experiments and analyzed data. A.F.D. and A.B.K. performed contact network analysis. A.F.D. prepared manuscript figures and wrote the first manuscript draft. R.C.T., J.C.F., A.B.K., M.M.B., K.E.P., F.C.P., and B.F.V. assisted in manuscript writing, interpreting the results, and revision. **Competing interests:** F.C.P. and B.F.V. have ownership interests in Protein Foundry, LLC. **Data and materials availability:** The NMR structure of Anc.O is available in the Protein Data Bank (PDB ID 7JH1) and the Biological Magnetic Resonance Data Bank (BMRB code 30777). Materials may be requested from the corresponding author.

## SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/371/6524/86/suppl/DC1  
Materials and Methods  
Supplementary Text  
Figs. S1 to S11  
Tables S1 to S3  
References (47–57)  
MDAR Reproducibility Checklist

20 July 2020; accepted 11 November 2020  
10.1126/science.abd8700

## ACTIVE MATTER

# Low rattling: A predictive principle for self-organization in active collectives

Pavel Chvykov<sup>1</sup>, Thomas A. Berrueta<sup>2</sup>, Akash Vardhan<sup>3</sup>, William Savoie<sup>3</sup>, Alexander Samland<sup>2</sup>, Todd D. Murphey<sup>2</sup>, Kurt Wiesenfeld<sup>3</sup>, Daniel I. Goldman<sup>3</sup>, Jeremy L. England<sup>3,4\*</sup>

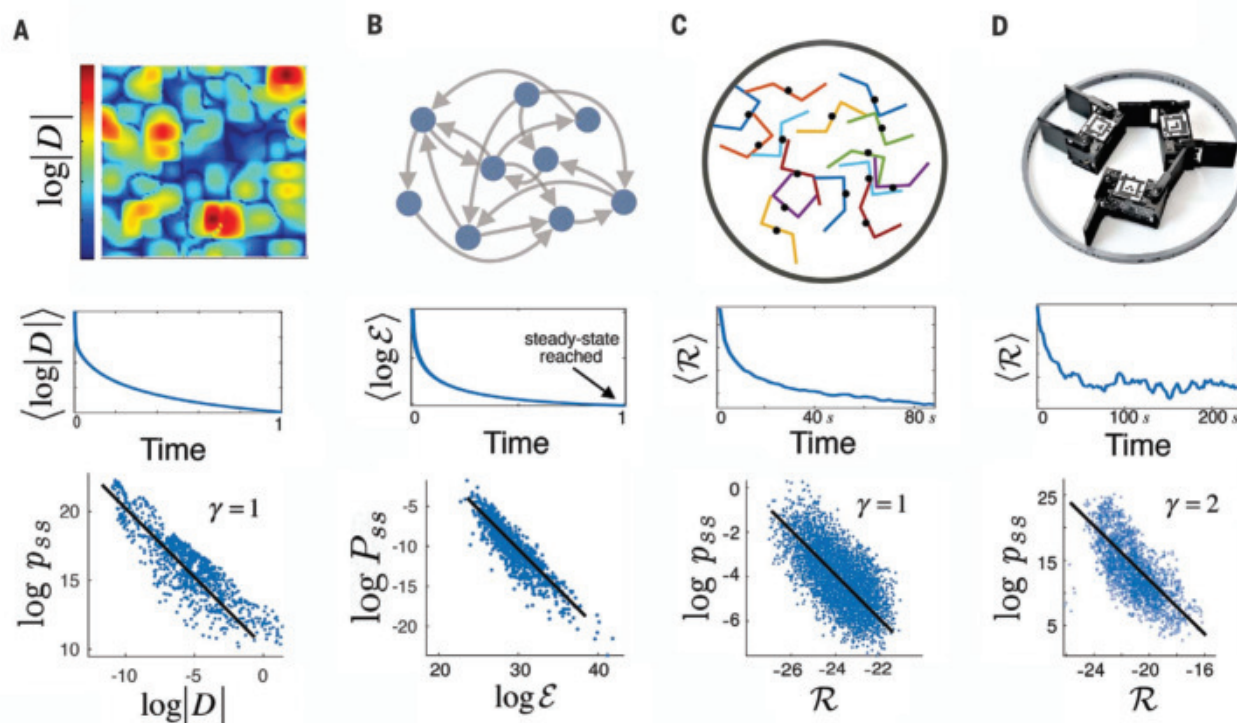
Self-organization is frequently observed in active collectives as varied as ant rafts and molecular motor assemblies. General principles describing self-organization away from equilibrium have been challenging to identify. We offer a unifying framework that models the behavior of complex systems as largely random while capturing their configuration-dependent response to external forcing. This allows derivation of a Boltzmann-like principle for understanding and manipulating driven self-organization. We validate our predictions experimentally, with the use of shape-changing robotic active matter, and outline a methodology for controlling collective behavior. Our findings highlight how emergent order depends sensitively on the matching between external patterns of forcing and internal dynamical response properties, pointing toward future approaches for the design and control of active particle mixtures and metamaterials.

Self-organization in nature is surprising because getting a large group of separate particles to act in an organized way is often difficult. By definition, arrangements of matter we call “orderly” are special, making up a tiny minority of all allowed configurations. For example, we find each unique, symmetrical shape of a snowflake visually striking, unlike any randomly rearranged clump of the same water molecules. Thus, any theory of emergent order in many-particle col-

lectives must explain how a small subset of configurations are spontaneously selected among the vast set of disorganized arrangements.

Spontaneous many-body order is well understood in thermal equilibrium cases such as crystalline solids or DNA origami (1), where the assembling matter is allowed to sit unperturbed for a long time at constant temperature  $T$ . The statistical mechanical approach proceeds by approximating the complex deterministic dynamics of the particles with a probabilistic “molecular chaos,” positing that the law of conservation of energy governs otherwise random behavior (2). What follows is the Boltzmann distribution for the steady-state probabilities,  $p_{ss}(\mathbf{q}) \propto \exp[-E(\mathbf{q})/T]$ , which shows that the degree to which special configurations  $\mathbf{q}$  of low energy  $E(\mathbf{q})$  have a high

<sup>1</sup>Physics of Living Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. <sup>2</sup>Department of Mechanical Engineering, Northwestern University, Evanston, IL 60208, USA. <sup>3</sup>School of Physics, Georgia Institute of Technology, Atlanta, GA 30332, USA. <sup>4</sup>GlaxoSmithKline AI/ML, 200 Cambridgepark Drive, Cambridge, MA 02140, USA. \*Corresponding author. Email: j@englandlab.com



**Fig. 1. Rattling  $\mathcal{R}$  is predictive of steady-state likelihood across far-from-equilibrium systems.** (A) Inhomogeneous anisotropic diffusion in two dimensions, where the steady-state density  $p_{ss}(\mathbf{q})$  is seen to be approximately given by the magnitude of local fluctuations  $\log|\mathbf{D}(\mathbf{q})| \propto \mathcal{R}(\mathbf{q})$  (where  $|\mathbf{D}|$  is the determinant of the diffusion tensor). (B) A random walk on a large random graph (1000 states), where  $P_{ss}$ , the probability at a state, is approximately given by  $\mathcal{E}$ , that state's exit rate. (C) An active matter system of shape-changing agents: an enclosed ensemble of 15 “smarticles” in simulation. (D) Experimental realization of similar agents with an enclosed three-robot smarticle ensemble. The middle row shows that relaxation to the

steady state of a uniform initial distribution is accompanied by monotonic decay in the average rattling value in all cases, analogous to free energy in equilibrium systems. The bottom row shows the validity of the nonequilibrium Boltzmann-like principle in Eq. 3, where the black lines in (A), (B), and (C) illustrate the theoretical correlation slope for a sufficiently large and complex system (see supplementary materials). The mesoscopic regime in (D) provides the most stringent test of rattling theory (where we observe deviations in  $\gamma$  from 1), while also exhibiting global self-organization. In the middle row, time units are arbitrary in (A) and (B); time is in seconds in (C) and (D), where the drive period is 2 s.

probability  $p_{ss}(\mathbf{q})$  in the long term depends on the amplitude of the thermal noise. Ordered configurations can assemble and remain stable, so long as interparticle attractions are strong enough to overcome the randomizing effects of thermal fluctuations.

However, there are also many examples of emergent order outside of thermal equilibrium. These include “random organization” in sheared colloids (3), phase separation in multitemperature particle mixtures (4), and dynamic vortices in protein filaments (5). A variety of ordered behaviors arise far from equilibrium that cannot be explained in terms of simple interparticle attraction or energy gradients (6–9).

In all of these examples, the energy flux from external sources allows different system configurations to experience fluctuations of different magnitude (10, 11). We suggest that the emergence of such configuration-dependent fluctuations, which cannot happen in equilibrium, may be key to understanding many nonequilibrium self-organization phenomena. In particular, we introduce a measure of

driving-induced random fluctuations, which we term rattling  $\mathcal{R}(\mathbf{q})$ , and argue that it could play a role in many far-from-equilibrium systems similar to the role of energy in equilibrium. We test this in a number of systems, including a flexible active matter system of simple robots we call “smarticles” (smart active particles) (12) as a convenient test platform (see movie S1) inspired by similar robo-physical emulators of collective behavior (13–15). Despite their purely repulsive inter-robot interactions, we find that smarticles spontaneously self-organize into collective “dances,” whose shape and motions are matched to the temporal pattern of external driving forces (movies S2 and S3). This platform and others (16–18), including the nonequilibrium ordering examples mentioned above, all exhibit low-rattling ordered behaviors that echo low-energy structures emergent at equilibrium. We thus motivate and test a predictive theory based on rattling that may explain a broad class of nonequilibrium ordering phenomena.

In devising our approach, we take inspiration from the phenomenon of thermophoresis,

which is the simplest example of purely nonequilibrium self-organization. Thermophoresis is characterized by the diffusion of colloidal particles from hot regions to cold regions (19). If noninteracting particles in a viscous fluid are subject to a temperature  $T(\mathbf{q})$  that varies over position  $\mathbf{q}$ , their resulting density in the steady-state  $p_{ss}(\mathbf{q})$  will concentrate in the regions of low temperature. Particles diffuse to regions where thermal noise is weaker, and they become trapped there. With the diffusivity landscape set by thermal noise locally according to the fluctuation-dissipation relation  $D(\mathbf{q}) \propto T(\mathbf{q})$  (20), the steady-state diffusion equation  $\nabla^2[D(\mathbf{q})p_{ss}(\mathbf{q})] = 0$  is satisfied by the probability density  $p_{ss}(\mathbf{q}) \propto 1/D(\mathbf{q})$ . Hence, a low-entropy, “ordered” arrangement of particles can be stable when the diffusivity landscape has a few locations  $\mathbf{q}$  that are strongly selected by their extremely low  $D(\mathbf{q})$  values.

We seek to extend this intuition to explain nonequilibrium self-organization more broadly. However, a straightforward mathematical extension of the idea encounters challenges in

only slightly more complicated scenarios. For an arbitrary diffusion tensor landscape  $\mathbf{D}(\mathbf{q})$ , in which diffusivity can depend on the direction of motion, one can no longer find general solutions for the steady state. Moreover, the steady-state density  $p_{ss}(\mathbf{q})$  at configuration  $\mathbf{q}$  may depend on the diffusivity  $\mathbf{D}(\tilde{\mathbf{q}})$  at arbitrarily distant configurations  $\tilde{\mathbf{q}}$ . Nonetheless, we suggest that for most typical diffusion landscapes, the local magnitude of fluctuations  $|\mathbf{D}(\mathbf{q})|$  should statistically bias  $p_{ss}(\mathbf{q})$  and hence should be approximately predictive of it. This insight, which is central to our theory, is illustrated to hold numerically in Fig. 1A for a randomly constructed two-dimensional anisotropic landscape, and in fig. S3 for higher dimensions. Although contrived counterexamples that break the relationship may be constructed, they require specific fine-tuning (see fig. S4).

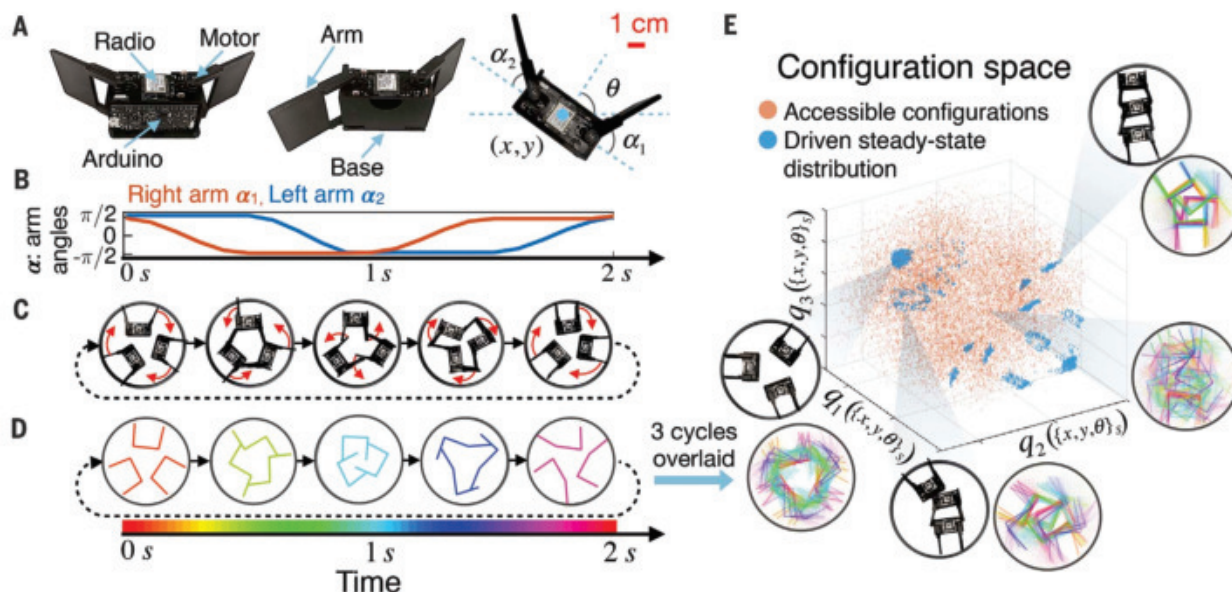
The key assumption underlying our approach is that the complex system dynamics are so messy that only the amplitude of local drive-induced fluctuations governs the otherwise random behavior—an assumption inspired by molecular chaos at equilibrium. We expect this to apply when the system dynamics are so complex, nonlinear, and high-dimensional that no global symmetry or constraint can be found for its simplification. Although one

cannot predict a configuration's nonequilibrium steady-state probability from its local properties in the general case (21, 22), the feat becomes achievable in practice for “messy” systems. To illustrate this point explicitly, we consider a discrete dynamical system with random transition rates between a large number of states. Here, we can show analytically that the net rate at which we exit any given state predicts its long-term probability approximately, even though the exact result requires global system knowledge (see Fig. 1B and supplementary materials for derivation). This result may be related to the above discussion of thermophoresis by noting that the discrete state exit rates are determined by the continuum diffusivity if our dynamics are built by discretizing the domain of a diffusion process.

To formulate our random dynamics assumption explicitly, we represent the complex system evolution as a trajectory in time  $\mathbf{q}(t)$ , where the configuration vector  $\mathbf{q}$  captures the properties of the entire many-particle system. Our messiness assumption amounts to approximating the full complex dynamics between two points  $\mathbf{q}(t)$  and  $\mathbf{q}(t + \delta t)$  by a random diffusion process. To this end, we take the amplitude of the noise fluctuations  $D(\mathbf{q})$  to locally reflect the amplitude of the true configuration dynamics:  $|\mathbf{q}(t + \delta t) - \mathbf{q}(t)|^2 \propto D(\mathbf{q})\delta t$  for short

rollouts  $\mathbf{q}(t \rightarrow t + \delta t)$  (i.e., samples of system trajectories) of duration  $\delta t$  initialized in configuration  $\mathbf{q}(t) = \mathbf{q}$  (see supplementary materials for details). Through this approximation, our dynamics are effectively reduced to diffusion in  $\mathbf{q}$ -space, which then allows us to locally estimate the steady-state probability of system configurations from  $D(\mathbf{q})$  as in thermophoresis. Hence, the global steady-state distribution may be predicted from the properties of short-time, local system rollouts.

For rare orderly configurations to be strongly selected in a messy dynamical system, the landscape of local fluctuations must vary in magnitude over a large range of values. Whereas in thermophoresis these fluctuations are directly imposed by an external temperature profile, in driven dynamical systems the range of magnitudes results from the way a given pattern of driving can have a different effect on different system configurations. The  $D(\mathbf{q})$  landscape is emergent from the interplay between the pattern of driving and the library of possible  $\mathbf{q}$ -dependent system response properties. In practice, we observe that the amplitudes of system responses to driving do often vary over several orders of magnitude (Fig. 1). We see this phenomenology in many well-known examples of active matter self-organization (3, 11, 23). For example, the crystals that form



**Fig. 2. Self-organization in a smarticle robotic ensemble.** (A) Front, back, and top views of a single smarticle. Of its five degrees of freedom, we consider the time-varying arm angles ( $\alpha_1, \alpha_2$ ) as “external” driving, because these are controlled by a preprogrammed microcontroller, whereas the robot coordinates ( $x, y, \theta$ ) are seen as an “internal” system configuration, because these respond interdependently to the arms. (B) An example of a periodic arm motion pattern. (C) Top view of three smarticles confined in a fixed ring, all programmed to synchronously execute the driving pattern shown in (B). The video frames, aligned on the time axis of (B), show one example of dynamically ordered collective “dance” that can spontaneously

emerge under this drive [see (E) and movie S3 for others]. (D) Simulation video showing agreement with experiment in (C). We color-code simulated states periodically in time and overlay them for three periods to illustrate the dynamical order over time. (E) The system's configuration space, built from nonlinear functions of the three robots' body coordinates ( $x, y, \theta$ ). The steady-state distribution (blue) illustrates the few ordered configurations that are spontaneously selected by the driving out of all accessible system states (orange). Simulation data are shown; see fig. S5B for experimental data and fig. S1 for details of how the configuration space coordinates ( $q_1, q_2, q_3$ ) in (E) are constructed from the  $3 \times (x, y, \theta)$  coordinates.



in suspensions of self-propelled colloids in (24) may be seen as the collective configurations that respond least diffusively to driving by precisely balancing the propulsive forces among individual particles. This illustrates how the low- $D(\mathbf{q})$  configurations are selected in the steady state by an exceptional matching of their response properties to the way the system is driven.

We apply these ideas in real complex driven systems whose response to driving we cannot predict analytically, such as our robotic swarm of smarticles. In this case, we require an estimator for the local value of  $D(\mathbf{q})$  based on observations of short rollouts of system behavior. The estimator of the local diffusion tensor that we choose here is the covariance matrix

$$\mathcal{C}(\mathbf{q}) = \text{cov}[\tilde{\mathbf{v}}_{\mathbf{q}}, \tilde{\mathbf{v}}_{\mathbf{q}}] \quad (1)$$

(25), where  $\tilde{\mathbf{v}}_{\mathbf{q}}$  is seen as a random variable with samples drawn from  $\{(\tilde{\mathbf{q}}(t) - \tilde{\mathbf{q}}(0))/\sqrt{t}\}_{\mathbf{q}(0)=\mathbf{q}}$  at various time points  $t$  along one or several short system trajectories  $\tilde{\mathbf{q}}(t)$  rolled out from  $\tilde{\mathbf{q}}(0) = \mathbf{q}$ . We assume these rollouts  $\tilde{\mathbf{q}}(t)$  to be long enough to capture fluctuations in the configuration variables under the influence of a drive, but short enough to have  $\tilde{\mathbf{q}}(t)$  stay near  $\mathbf{q}$  (see supplementary materials for details).

Although the covariance matrix reflects the amplitude of local fluctuations, we are instead interested in a measure of their disorder if we want to estimate the effective diffusivity. This follows from the observation that high-amplitude ordered oscillations do not contribute to the rate of stochastic diffusion (10).

We suggest that the degree of disorder of fluctuations may be captured by the entropy of the distribution of  $\tilde{\mathbf{v}}_{\mathbf{q}}$  vectors, which is how we define rattling  $\mathcal{R}(\mathbf{q})$ . Physically, vectors  $\tilde{\mathbf{v}}_{\mathbf{q}}$  capture the statistics of the force fluctuations experienced in configuration  $\mathbf{q}$ , and so rattling measures the disorder in the system's driven response properties at that point. By approximating the distribution of  $\tilde{\mathbf{v}}_{\mathbf{q}}$  as Gaussian, we can express its entropy (up to a constant offset) simply in terms of  $\mathcal{C}(\mathbf{q})$  as

$$\mathcal{R}(\mathbf{q}) = \frac{1}{2} \log \det \mathcal{C}(\mathbf{q}) \quad (2)$$

With this definition, we generalize the thermophoretic expression for the steady-state density  $p_{ss}(\mathbf{q}) \propto 1/D(\mathbf{q})$  and express it in a Boltzmann-like form:

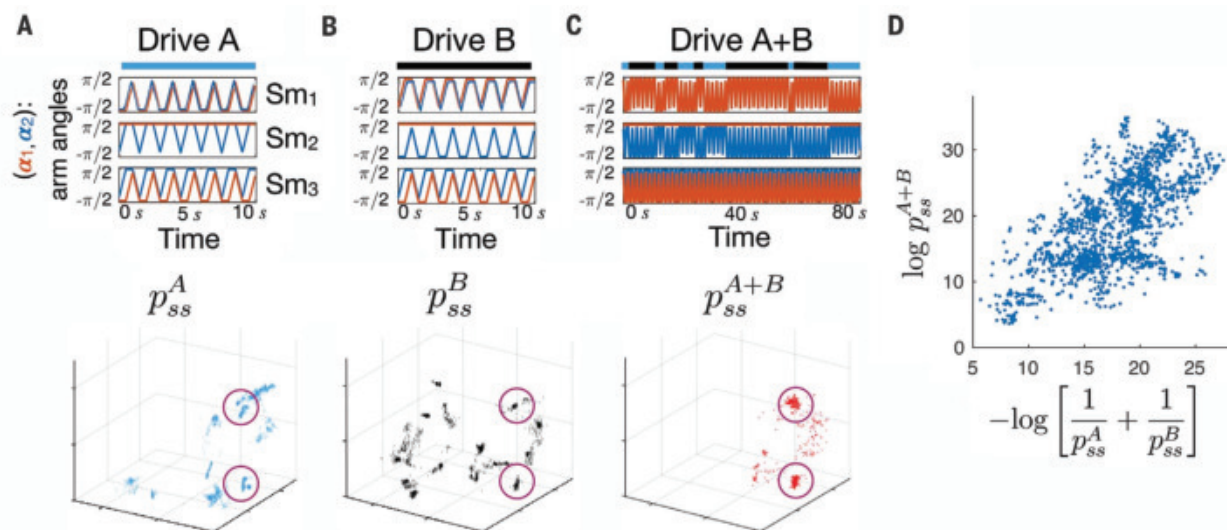
$$p_{ss}(\mathbf{q}) \propto \exp[-\gamma \mathcal{R}(\mathbf{q})] \quad (3)$$

where  $\gamma$  is a system-specific constant of order 1 (see supplementary materials for derivations). We note that when energy varies on the same scale as rattling, the interaction between the two landscapes can generate strong steady-state currents and may break this relation (10). Thus, rattling enables us to predict the long-term global steady-state distribution based on empirical measurements of short-term local system behavior, which suggests that probability density accumulates over time in low-rattling configurations.

We study the collective behavior of a simple ensemble of smarticles, aligning ourselves within the tradition of using robotic systems as flexible, physical emulators for self-organizing natural systems (13–16). Each smarticle (Fig. 2A)

is composed of three 5.2-cm links, with two hinges actuated by motors programmed to follow a driving pattern specified by a microcontroller. When a smarticle sits on a flat surface, its arms do not touch the ground, so an individual robot cannot move. However, a group of them can achieve complex motion by pushing and pulling each other (movie S1) (26). The relative coordinates of the middle link of each robot in the ensemble  $(x, y, \theta)$  may be thought of as the internal system configurations that dynamically respond to an externally determined driving force arising from the time variation of arm angles  $(\alpha_1, \alpha_2)$  (27).

This robotic active matter system offers substantial flexibility in choosing the programmed patterns of driving as well as the properties of internal system dynamics (friction coefficients, weights, etc.). Additionally, the smarticle system has a flat potential energy landscape, allowing one to focus on the contributions of the drive-induced fluctuations to the collective behavior, which makes our findings broadly applicable to other strongly driven systems. When the smarticles are within contact range (as ensured by a confining ring; Fig. 1D), the forces experienced throughout the collective for a given pattern of arm movement are an emergent function of all system coordinates. This configuration-dependent forcing gives rise to varying rattling values, which we refer to as the “rattling landscape,” and which we see to be a hallmark property in many far-from-equilibrium examples. The rattling landscape then leads to some system configurations being dynamically selected over others and allowing for self-organization, just as the diffusivity landscape does in thermophoresis. Finally,



**Fig. 3. Self-organized behaviors are fine-tuned to drive pattern.** (A and B) Changing the arm motion pattern slightly (top) affects which configurations self-organize in the steady state (bottom, same 3D configuration space as in Fig. 2E). (C) By mixing drives A and B as shown (top), we can isolate only those

configurations selected in both the steady states (circled in purple; see movie S6), which follows as an analytical prediction of the theory. (D) This prediction (Eq. 4) is quantitatively verified. All data shown are experimental and are reproduced in simulation in fig. S7, along with derivations in the supplementary materials.

the combined effects of impulsive inter-robot collisions, nonlinear boundary interactions, and static friction lead to a large degree of quasi-random motion (26), making this a promising candidate system for exploring our theory.

Reasoning that our fundamental assumption of quasi-random configuration dynamics

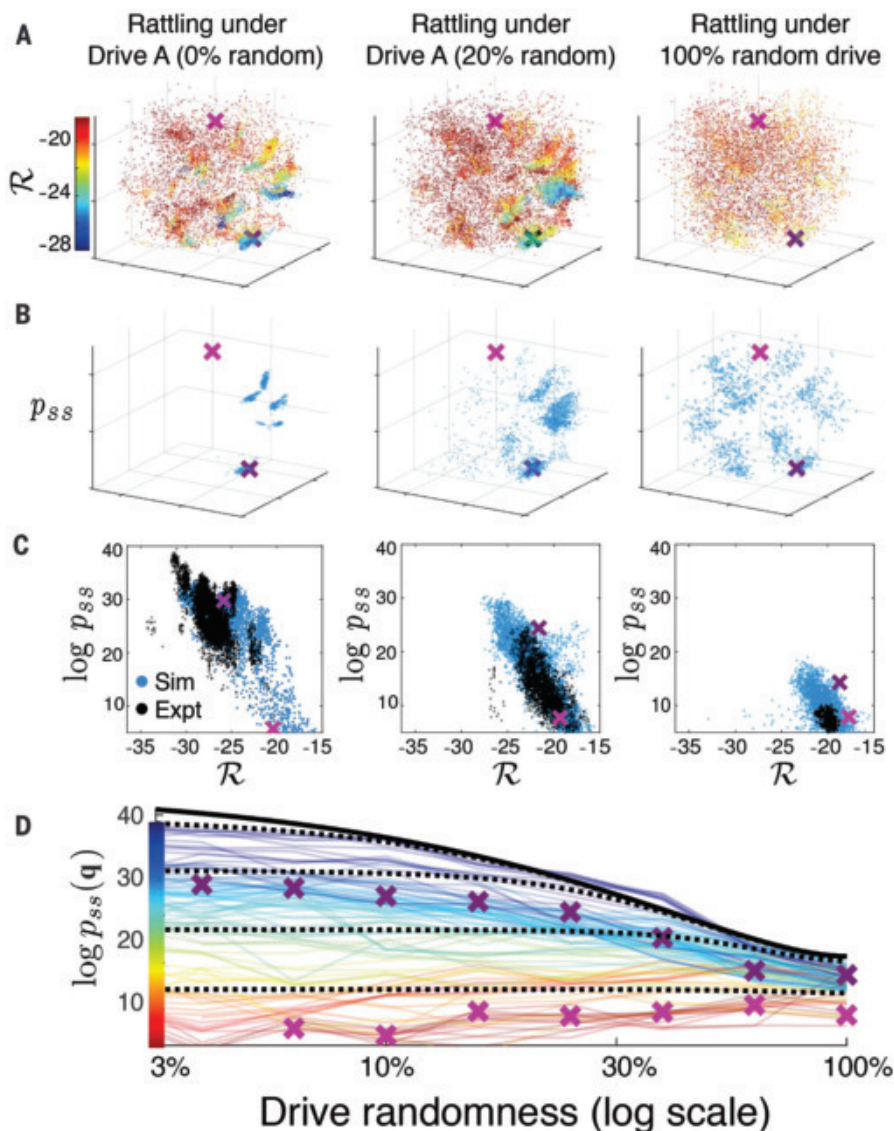
would be most valid in systems with many degrees of freedom, we also built a simulation that would allow us to study the properties of larger smarticle groups and explore different system parameters (fig. S9). In this regime, we used simulations to gather enough data to sample the high-dimensional probability distributions for our analysis. In a simulation of

15 smarticles, we observed the tendency of the ensemble to reduce average rattling over time after a random initialization. For this 45-dimensional system ( $x, y, \theta$  for 15 robots), the configuration-space dynamics are well approximated by diffusion, and so Eq. 3 holds, as seen in Fig. 1C. In addition, we noted the emergence of metastable pockets of local order when groups of three or four nearby smarticles self-organized into regular motion patterns for several drive cycles (movie S2). A signature of such dynamical heterogeneity can be seen in the spectrum of the covariance matrix  $C(\mathbf{q})$  from Eq. 1, as described in the supplementary materials and fig. S10.

The transient appearance of dynamical order in subsets of smarticle collectives raises the question of whether our rattling theory continues to hold for smaller ensembles. For the remainder of this paper, we focus on ensembles of three smarticles (as in Fig. 1D), which allows for exhaustive sampling of configurations experimentally, as well as easier visualization of the configuration space (as in Fig. 2E). Both in simulation and experiment, we found that this regime exhibits a variety of low-rattling behaviors that manifest as distinct, orderly collective “dances” (Fig. 2, C and D, and movie S3). Despite its small size, this system is well described by rattling theory, as evidenced by the empirical correlation between rattling and the steady-state likelihood of configurations (Fig. 1D, bottom).

We consider self-organization as a consequence of a system's landscape of rattling values over configuration space. This rattling landscape is specific to the particular drive forcing the system out of equilibrium, because different drives will generally produce different dynamical responses in the same system configuration. When the three-smarticle ensemble is driven (under the pattern in Fig. 2B), the range of observed rattling values is so large that the lowest-rattling configurations—and consequently those with the highest likelihood—account for most of the steady-state probability mass. More than 99% of probability accumulates in these spontaneously selected configurations, which represent only 0.1% of all accessible system states (Fig. 2E). Moreover, in these configurations the smarticles exhibit an orderly response to driving (Fig. 2, C and D, and movie S4). In practice, the ensemble spends most of its time in or nearly in one of several distinct dances, with occasional interruptions by stochastic flights from one such dynamical attractor to another (movie S5).

From the above observations, we can begin to understand self-organization in driven collectives. In equilibrium, order arises when its entropic cost is outweighed by the available reduction of energy. Analogously, a sufficiently large reduction in rattling can lead to dynamical organization in a driven system.



**Fig. 4. Tuning self-organization by modulating drive randomness.** Self-organization relies on the degree of predictability in its driving forces, in a way that we can quantify and compute analytically. (A) As the drive becomes less predictable (left to right in all panels), low-rattling configurations gradually disappear. (B) The corresponding steady states, reflecting the low-rattling regions of (A), become accordingly more diffuse. [(A) and (B) show simulation data and use the same 3D configuration space as Fig. 2E]. (C) All three correlations fall along the same line (blue, simulation; black, experiment), verifying that our central predictive relation (Eq. 3) holds for all drives here. The diminishing range of rattling values thus precludes strong aggregation of probability, and with it self-organization. (D) Our theoretical prediction (solid black line) indicating how the most likely configurations are destabilized by drive randomness. Colored lines track the probability  $p_{ss}$  at 100 representative configurations  $\mathbf{q}$  in simulation, and dashed black lines analytically predict their trends. (movie S8; see supplementary materials for derivation). Two specific configurations marked by pink and purple crosses are tracked across analyses.



Moreover, such a reduction can require matching between the system dynamics and the drive pattern.

Through rattling theory we can predict how self-organized states are affected by changes in the features of the drive. We expect the structure of the self-organized dynamical attractors to be specific to the driving pattern, as each drive induces its own rattling landscape. To test this, we programmed the three smarticles with two distinct driving patterns (Fig. 3, A and B, top), which we ran separately. The two resulting steady-state distributions, although each is highly localized to a few configurations, are largely non-overlapping (Fig. 3, A and B, bottom). This indicates that by tuning the drive pattern, it may be possible to design the structure of the resulting steady state, and hence to control the self-organized dynamics [see also (28–30)].

As a proof of principle for such control, we developed a methodology for selecting particular steady-state behaviors by combining drives. By randomly switching back and forth between drives A and B in Fig. 3, we define a compound drive A+B (Fig. 3C and movie S6). We predicted that this drive would select only those configurations common to both A and B steady states (Fig. 3, A and B, bottom), because having low rattling under this mixed drive requires having low rattling under both constituent drives. Our experiments confirmed this (Fig. 3C), and we were further able to quantitatively predict the probability that a configuration would appear under the mixed drive on the basis of its likelihood in each constituent steady state according to

$$\frac{1}{p_{ss}^{A+B}} \propto \frac{1}{p_{ss}^A} + \frac{1}{p_{ss}^B} \quad (4)$$

as shown in Fig. 3D and fig. S7 (see supplementary materials for derivation). This simple relationship suggests that by composing different drives in time, one can single out desired configurations for the system steady state.

Moreover, we show that we can analytically predict and control the degree of order in the system by tuning drive randomness (Fig. 4) as well as internal system friction (movie S7, fig. S8, and supplementary materials). Because driven self-organization arises when the system has access to a broad range of rattling values, tuning it requires modulating the rattling of the most ordered behaviors relative to the background high-rattling states.

We can directly manipulate the rattling landscape by modulating the entropy of the drive pattern. This is done by introducing a probabilistic element to the programmed arm motion. At each move, we introduce a probability of making a random arm movement not included in the prescribed drive pattern. In-

creasing this probability results in flattening the rattling landscape: Ordered states experience an increase in rattling due to drive entropy, whereas states whose rattling is already high do not (Fig. 4A). Correspondingly, the steady-state distributions become progressively more diffuse (Fig. 4B), causing localized pockets of order to give way to entropy and “melt” away—just as crystals might in equilibrium physics [movie S8; see also (37)].

Even as the range of accessible rattling values in the system shrinks, the predictive relation of Eq. 3 continues to hold (Fig. 4C), enabling quantitative prediction of how self-organized configurations are destabilized. By calculating the entropy of the drive pattern as we tune its randomness, we derive a lower bound on rattling for the system. Thus, we can analytically predict how steady-state probabilities change as a function of drive randomness, as shown in Fig. 4D (up to normalization and  $\gamma$ ; see supplementary materials for derivation). This result confirms the simple intuition that more predictably patterned driving forces offer greater opportunity for the system to find low-rattling configurations and self-organize (see also fig. S6).

Our findings suggest that the complex dynamics of a driven collective of nonlinearly interacting particles may give rise to a situation in which a new kind of simplicity emerges. We have shown that when quasi-random transitions among configurations dominate the dynamics, the steady-state likelihood can be predicted from the entropy of local force fluctuations, which we refer to as rattling. In what we term a “low-rattling selection principle,” configurations are selected in the steady state according to their rattling values under a given drive.

Low rattling provides the basis for self-organized dynamical order that is specifically selected by the choice of driving pattern. We see analytically and experimentally that the degree of order in the steady-state distribution reflects the predictability of patterns in driving forces. Thus, driving patterns with low entropy pick out fine-tuned configurations and dynamical trajectories to stabilize. This makes it possible for one collective to exhibit different modes of ordered motion depending on the fingerprint of the external driving. These modes differ in their emergent collective properties, which suggests “top-down” alternatives to control of active matter and metamaterial design, where ensemble behaviors, rather than being microscopically engineered, are dynamically self-selected by the choice of driving (30, 32).

## REFERENCES AND NOTES

1. P. W. K. Rothmund, *Nature* **440**, 297–302 (2006).
2. M. Kardar, *Statistical Physics of Particles* (Cambridge Univ. Press, 2007).

3. L. Corté, P. Chaikin, J. P. Gollub, D. Pine, *Nat. Phys.* **4**, 420–424 (2008).
4. A. Y. Grosberg, J.-F. Joanny, *Phys. Rev. E* **92**, 032118 (2015).
5. Y. Sumino *et al.*, *Nature* **483**, 448–452 (2012).
6. S. Ramaswamy, *Annu. Rev. Condens. Matter Phys.* **1**, 323–345 (2010).
7. L. Bertini, A. De Sole, D. Gabrielli, G. Jona-Lasinio, C. Landim, *Rev. Mod. Phys.* **87**, 593–636 (2015).
8. M. Paoluzzi, C. Maggi, U. Marini Bettolo Marconi, N. Gnan, *Phys. Rev. E* **94**, 052602 (2016).
9. T. Speck, *Europhys. Lett.* **114**, 30006 (2016).
10. P. Chvykov, J. England, *Phys. Rev. E* **97**, 032115 (2018).
11. M. E. Cates, J. Tailleur, *Annu. Rev. Condens. Matter Phys.* **6**, 219–244 (2015).
12. W. Savoie, thesis, Georgia Institute of Technology (2019).
13. J. Aguilar *et al.*, *Science* **361**, 672–677 (2018).
14. M. Rubenstein, A. Cornejo, R. Nagpal, *Science* **345**, 795–799 (2014).
15. J. Werfel, K. Petersen, R. Nagpal, *Science* **343**, 754–758 (2014).
16. S. Li *et al.*, *Nature* **567**, 361–365 (2019).
17. G. Várhelyi *et al.*, *Sci. Robot.* **3**, eaat3536 (2018).
18. S. Mayya, G. Notomista, D. Shell, S. Hutchinson, M. Egerstedt, in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Macau, China (2019), pp. 4106–4112.
19. S. Dühr, D. Braun, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 19678–19682 (2006).
20. N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry*, vol. 1 (Elsevier, 1992).
21. R. Landauer, *Physica A* **194**, 551–562 (1993).
22. R. Landauer, *Phys. Rev. A* **12**, 636–638 (1975).
23. G. S. Redner, M. F. Hagan, A. Baskaran, *Phys. Rev. Lett.* **110**, 055701 (2013).
24. J. Palacci, S. Sacanna, A. P. Steinberg, D. J. Pine, P. M. Chaikin, *Science* **339**, 936–940 (2013).
25. X. Michalet, A. J. Berglund, *Phys. Rev. E* **85**, 061916 (2012).
26. W. Savoie *et al.*, *Sci. Robot.* **4**, eaax4316 (2019).
27. See supplementary materials.
28. H. Kedia, D. Pan, J.-J. Slotine, J. L. England, arXiv 1908.09332 [nlin.AO] (25 August 2019).
29. T. Epstein, J. Fineberg, *Phys. Rev. Lett.* **92**, 244502 (2004).
30. H. Karani, G. E. Pradillo, P. M. Vlahovska, *Phys. Rev. Lett.* **123**, 208002 (2019).
31. D. I. Goldman, M. D. Shattuck, S. J. Moon, J. B. Swift, H. L. Swinney, *Phys. Rev. Lett.* **90**, 104302 (2003).
32. O. Sigmund, in *IUTAM Symposium on Modelling Nanomaterials and Nanosystems* (Springer, 2009), pp. 151–159.
33. T. A. Berrueta, A. Samland, P. Chvykov, Repository with all data, hardware, firmware, and software files needed to recreate the results of this manuscript: <https://doi.org/10.5281/zenodo.4056700>.

## ACKNOWLEDGMENTS

We thank P. Umbanhowar, H. Kedia, and J. Owen for helpful discussions. **Funding:** Supported by ARO grant W911NF-18-1-0101 and James S. McDonnell Foundation Scholar grant 220020476 (P.C. and J.L.E.); ARO MURI award W911NF-19-1-0233 and NSF grant CBET-1637764 (T.A.B., A.S., and T.D.M.); NSF grants PoLS-0957659, PHY-1205878, and DMR-1551095 and ARO grant W911NF-13-1-0347 (A.V., W.S., and D.I.G.); and NSF grant PHY-1205878 (K.W.). **Author contributions:** P.C. derived all theoretical results, performed simulations, data analysis, and contributed to writing; T.A.B. performed all experiments in the main text and contributed to writing and data analysis; A.V. and W.S. performed supplementary experiments; A.S. aided in robot hardware and software fabrication; and J.L.E., D.I.G., K.W., and T.D.M. secured funding and provided guidance throughout. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All files needed for fabricating smarticles, as well as representative data, can be found in (33).

## SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/371/6524/90/suppl/DC1  
Materials and Methods  
Supplementary Text  
Figs. S1 to S10  
References (34–46)  
Movies S1 to S8

6 May 2020; accepted 27 November 2020  
10.1126/science.abc6182



# FIND YOUR HAPPIER PLACE.



Find your next job at [ScienceCareers.org](https://www.sciencecareers.org)

There's scientific proof that when you're happy with what you do, you're better at what you do. Access career opportunities, see who's hiring and take advantage of our proprietary career-search tools. Get tailored job alerts, post your resume and manage your applications all in one place: [sciencecareers.org](https://www.sciencecareers.org)

**ScienceCareers**

FROM THE JOURNAL SCIENCE  AAAS

# Who's the top employer for 2020?

*Science Careers'* annual survey reveals the top companies in biotech & pharma voted on by *Science* readers.

Read the article and employer profiles at [sciencecareers.org/topemployers](https://sciencecareers.org/topemployers)



**Science 2020  
TOP EMPLOYER**



By Angela Q. Zhang

## Saying yes to help

**“A**re you OK?” my principal investigator (PI) asked me. I had just broken down crying in his office during one of our meetings. “It doesn’t seem like you’re OK.” He was right. But I wasn’t ready to be vulnerable with him, so I evaded the question. Later, I wondered why. When I mentor undergrads, I make a point of connecting with them on a personal level and reaching out to them when they seem to need help. For the past year, I had been yearning for someone to do the same for me. So why hadn’t I accepted the gesture when it finally came?

Things had started to go downhill for me during the third year of my Ph.D. My science wasn’t going as planned, and I was in the midst of a long-standing conflict with a colleague. I was dragging myself into the lab at 1 p.m., my face hidden beneath my hood, headphones on to drown out the chatter around me. I stopped speaking up in meetings. The quality and quantity of my work dropped. With my downcast eyes, slow gait, and slumped posture, I tried to signal that I needed help—but nobody reached out. To my labmates, I may have just seemed stressed or tired. As for my PI, he seemed to not want to pry into my personal life. I felt alone and helpless, hesitant to share my struggles because I wasn’t sure that anyone cared.

For a while, my undergrads kept me functioning. Their curiosity spurred me to plan experiments and read papers. My duty to them forced me out of bed and into the lab, where I set aside my own distress and put on the disguise of an encouraging mentor. I enthusiastically asked about their classes and weekend plans, their extracurricular activities and postgraduation ambitions. Mentoring offered a consolation: If I couldn’t make my mark in science, at least I could have an impact on my mentees’ career trajectories and support them through their own challenges.

When one of my undergrads began to act lethargic and distracted, for example, I reached out to ask whether there was anything I could do. Though usually reticent, she opened up. She thanked me for checking in and offering a sympathetic ear, and we adjusted her lab workload to accommodate her needs. Why was it that I could be there for my mentees, yet no one could be there for me?

Soon enough, I lost the high I got from mentoring. My patience gave way to irritation. When my undergrads



**“Concealing my struggles ... hampered my ability to be my true self in the lab.”**

made simple mistakes, I had a harder time being understanding. I knew that I couldn’t wait any longer to seek help. I finally contacted the therapist I had connected with at the beginning of grad school and started medication for my now-diagnosed anxiety and depression.

Then came that meeting with my adviser. Because our relationship had always been strictly professional, I wasn’t sure he really wanted to know about my troubles any more than my other colleagues seemed to. I also worried that he would think less of me if I told him I was having a hard time.

Yet concealing my struggles from my closest colleagues hampered my ability to be my true self in the lab—the place where I spent most of my waking hours.

Eventually, I worked up the courage to tell close friends and supportive labmates. Many listened and empathized, and I realized that just because they hadn’t reached out didn’t mean they didn’t care.

In the end, I told my PI too. It was incredibly awkward at first, but with time, we became more comfortable having frank, candid conversations. We made arrangements to minimize the conflict I had with my colleague and devised a plan to balance my scientific interests with my graduation timeline. I’m still working on my mental health, but I finally feel like I’m headed in the right direction.

The experience has taught me that when I need help and support, sometimes I need to ask for it. And when it’s offered, even from an unlikely source, I should embrace it. ■

Angela Q. Zhang is an M.D.-Ph.D. candidate in the Harvard-MIT Program in Health Sciences and Technology. Send your career story to [SciCareerEditor@aaas.org](mailto:SciCareerEditor@aaas.org).