New hope for green energy from cheap, bendy films p. 588

NOMIS & *Science* Award for culture's influence on cognition p. 610

Drought shaped early plant vessel evolution p. 642

Science

$15
11 NOVEMBER 2022
science.org
AAAS



WIRES FROM DROPLETS

Using acoustics to create circuits pp. 594 & 637

EDITORIAL

Be the voice for scientists in Iran



Navid Madani is the founding director of the Science Health Education Center at the Dana-Farber Cancer Institute (DFCI), Boston, MA, USA. She is a lead scientist in the Department of Cancer Immunology and Virology at DFCI and an affiliate faculty member in the Departments of Microbiology and Global Health and Social Medicine at Harvard Medical School, Boston, MA, USA. navid_madani@dfci.harvard.edu

Iran's fundamentalist government has long feared students and academics, because independent thinking and inquiry are at odds with the extreme rhetoric of a repressive religious regime that discourages questioning or testing—especially when those asking questions are women. But when the Iranian "morality police" beat 22-year-old Mahsa Amini to death in September for wearing "un-Islamic clothing," they unintentionally restored the long-silent voices of Iranian students, scholars, and scientists.

In the 1980s, Ruhollah Khomeini, founder of the Islamic Republic of Iran, described the great danger that university-educated people posed to the regime. This philosophy was incongruous with Iran's centuries-long history as a beacon of science, philosophy, and medicine. In the 11th century, for example, the Iranian physician-philosopher Ibn Sina wrote about cancer metastasis for the first time in human history. Subsequent advances in medicine and science flourished in the Iranian academic landscape. Khomeini's words were a direct attack on intellectuals, leading me and other young scientists to leave Iran and pursue professional careers elsewhere.

Since the 1979 revolution, Iran's leadership has increasingly attacked science and critical thinking, making the country's brain drain among the highest in the world. Pressure to conform to rigid, strictly enforced behavioral constraints falls heaviest on young people, especially women. Those who speak against this are jailed, abused, and murdered—just as Mahsa Amini was. Despite these injustices, over 60% of Iranian women are college graduates. Iran gave the world mathematician Maryam Mirzakhani—the first woman to win the Fields Medal—and Nobel Peace Prize–winning activist Shirin Ebadi. Iran has thousands more students and young professionals eager to further their explorations of science and medicine.

The ongoing public outrage sparked by Amini's senseless murder reflects people tired of watching authorities end their dreams with arbitrary violence. Recently, a surgeon was shot and killed by security forces as her physician colleagues protested in front of the Islamic Republic Medical Council. Indeed, infuriated Iranians around the world are protesting Iran's oppression of women. Last month, members of the global academic community in North America condemned a brutal attack on students and faculty at Sharif University of Technology in Tehran.

As an Iranian female scientist in the United States, I believe that all scientists have a distinct duty to recognize these attacks as an assault on education, knowledge, and human rights; to speak out against injustices in the name of religion; and to give a voice to students, scientists, and health care professionals living in oppressive societies. After finishing my doctoral dissertation in 1999, I returned to Iran to give a scientific talk at an all-women's university north of Tehran. Most attendees were young women pursuing scientific training. Their eagerness and enthusiasm about scientific study convinced me that Iran was returning to its roots as a nation that valued inquiry. But dismayingly, Iran's students and scholars are again willing to risk jail, torture, and death to pursue knowledge. As I watch this uprising, I wonder if any of the young women I met in 1999 are part of this revolution in some way, supporting the next generation of brave women who are fighting for change.

"How can scientists across the free world...turn outrage into action?"

How can scientists across the free world show solidarity and turn outrage into action? There is power in numbers. The global scientific community can spread the word of this inhumanity by amplifying the voices of the people in Iran through traditional and social media and in scientific gatherings; by signing the Amnesty International petition calling for the United Nations Human Rights Council to hold Iran accountable for the violence; and by supporting the work of the Abdorrahman Boroumand Center for Human Rights in Iran, a nongovernmental and nonprofit organization that promotes democracy. Above all, scientists can help students in Iran continue to learn by providing them online resources if their universities cannot. As an example, my colleagues and I are creating a Persian scientific curricula for university students in subjects such as biochemistry and math so that they can study while classes are shut down because of protests.

Scientists largely have stood by in silence for decades as Iran's unique scientific heritage was denigrated. It's time to speak up.

–Navid Madani

10.1126/science.adf6588

PHOTO: SAM OGDEN/DANA FARBER CANCER INSTITUTE

Higher education for all



Marcia McNutt is president of the United States National Academy of Sciences, Washington, DC, USA. mmcnutt@nas.edu

Universities are one of the oldest human institutions, enduring with essentially the same blueprint for a thousand years. Before the current COVID-19 pandemic, there was much talk about the promise of massive open online courses, distance learning, and other innovations to scale and expand the reach of universities, but with only limited success. Given the experience gained from educating during the pandemic, it is time for educators to ask which innovations can be introduced and, importantly, sustained, to expand the accessibility of higher education to meet the needs of the 21st century.

Currently, 75% of new jobs require a college degree. Yet in the US and Europe, only 40% of young adults attend a 2-year or 4-year college—a percentage that has either not budged or only modestly risen in more than two decades—despite a college education being one of the proven ways to lift the socioeconomic status of underprivileged populations and boost the wealth of nations. Worse, only 18% of that 40% receive degrees in a STEM (science, technology, engineering, and mathematics) discipline, although the fraction has been slowly rising over the past decade. Depending on foreign STEM students from Asia to fill the gap is not a viable solution. In the near future, workers in low-skill jobs without college degrees are at risk of being replaced by automation. In fact, improved access to additional training will be a life-long need as the pace of innovation exceeds the duration of an individual career. And the complex challenges that we face as a global society, including climate change and unsustainable use of resources, demand a higher level of educational attainment for all, regardless of job description.

> **"...expansion... requires many young adults to pull up roots...while also working to support a family."**

A college education in the US is expensive, even at public institutions, and it can take many years, if ever, to recoup the cost through future earnings. Tuition, however, is not the main impediment. Enrollment of young adults is stagnant even in nations where tuition is free. The bigger barrier to expansion appears to be the traditional college residential program that requires many young adults to pull up roots and move to a new location to pursue a degree while also working to support a family.

Universities, without building additional facilities, could expand universal and life-long access to higher education by promoting more courses online and at satellite community-college campuses. Science education, with its need for collaboration, fieldwork, and hands-on exploration, presents special challenges. Nevertheless, solutions are at hand. At Colorado College, students complete a lab science course in only 4 weeks, attending lectures in the morning and labs in the afternoon. This success suggests that US universities could offer 2-week short courses that include concentrated, hands-on learning and teamwork in the lab and the field for students who already mastered the basics through online lectures. Such an approach is more common in European institutions of higher education and would allow even those with full-time employment elsewhere to advance their skills during vacations or employer-supported sabbaticals for the purpose of improving the skills of the workforce. Opportunities abound for partnerships with industry for life-long learning. The availability of science training in this format could also be a boon for teachers seeking to fill gaps in their science understanding.

State universities are leading in experimentation in new formats. Arizona State University has increased its engineering majors by a factor of 5 to more than 30,000 students in the past 10 years, including 8500 online learners. Lab skills are taught during 2-week summer sessions. The degree conferred to online students is indistinguishable from that earned by students completing a 4-year residential program. Ohio State University is providing more opportunities for education in sought-after technology fields by working with colleges across multiple states to reach students where they live.

A university education may not be for everyone, but at least we must make it more easily available to those who would benefit without the traditional obstacles to completing a 4-year residential program. By committing to educating citizens throughout their lifetimes in a manner that respects the realities of income, work, family, culture, and community, universities will reach their true potential as generators of the most precious of all resources: human capital, concentrated not just in university towns but throughout the country.

–Marcia McNutt

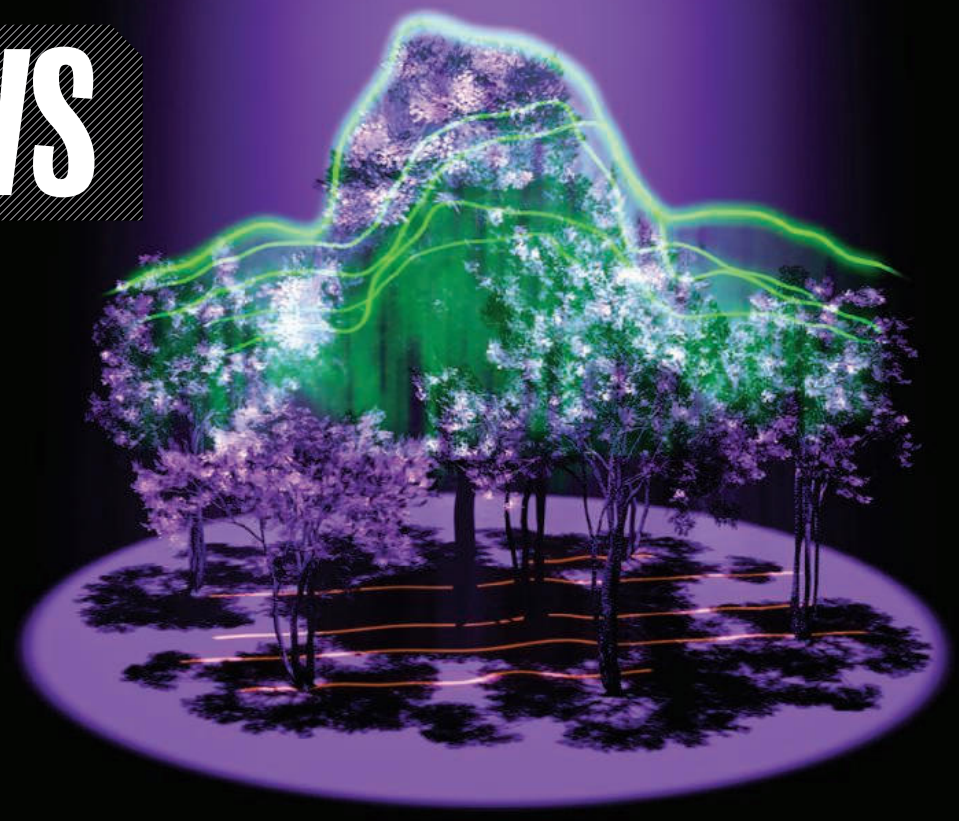
PHOTO: CHRIS MICHEL

10.1126/science.adf6263

NEWS

IN BRIEF

Edited by
Jeffrey Brainard



CLIMATE SCIENCE

Space station carbon mapper faces demise

An orbiting fridge-size sensor that uses lasers to map forest structure—key to understanding how much carbon trees sequester—is set to plummet to a fiery destruction in the atmosphere in 2023 unless NASA extends its tenure. Researchers and some U.S. Congress members are lobbying the agency to reconsider its plan to jettison the Global Ecosystem Dynamics Investigation (GEDI) instrument from the International Space Station to make way for a Department of Defense sensor. GEDI measures the height of trees and the quality of habitat they provide, information that its supporters say imaging satellites such as Landsat cannot provide. In operation since April 2019, GEDI has identified the parts of the Amazon rainforest that hold the most carbon, which could guide conservation efforts intended to minimize deforestation and the release of carbon to the atmosphere. GEDI's value has been constrained by a lack of ground data from locations such as China and Indonesia, which are necessary to calibrate its readings.

A space-based laser helps create 3D images of forests, improving estimates of how much carbon they store.

IMAGE: NASA GODDARD SPACE FLIGHT CENTER

New antibiotic passes key hurdle

DRUG DEVELOPMENT | A new antibiotic that represents an entirely novel chemical class has passed it first clinical test. The drug, gepotidacin, cured urinary tract infections (UTIs) so well in two large trials that researchers stopped them early. Its manufacturer, GSK, says it plans to seek approval of the drug from the U.S. Food and Drug Administration early next year; if it succeeds, gepotidacin will be the first new oral antibiotic for common UTIs in more than 20 years. Gepotidacin inhibits bacterial DNA replication by blocking two essential enzymes, gyrase and topoisomerase IV. This makes it effective against most strains of *Escherichia coli*, the primary culprit in UTIs, including those resistant to the fluoroquinolones, the current front-line antibiotics. Gepotidacin is also being investigated as a treatment for gonorrhea. Bacteria would likely need mutations in both targeted enzymes to dodge the drug, raising hopes that resistance won't develop easily.

COVID-19 hits Antarctic station

POLAR SCIENCE | McMurdo Station, the largest research outpost in Antarctica, is suffering from an unprecedented outbreak of COVID-19, with at least 64 active cases among its more than 900 residents, the U.S. National Science Foundation (NSF) said on 7 November. The agency this week paused most flights to the continent for 2 weeks and recommended that all residents wear KN-95 masks. This year, NSF relaxed its strict policies that had required researchers to isolate before traveling to Antarctica, but the agency's vaccine mandate, which now requires a bivalent booster, remains in force. The outbreak will likely upset summer fieldwork, much of which relies on McMurdo as a logistical hub, further straining projects that the pandemic has already delayed by several years.

Novel defenses help infant lungs

BIOMEDICINE | The world may soon have two new defenses against respiratory syncytial virus (RSV), a potentially fatal pathogen that has filled U.S. pediatric

> “The damage is obvious. Those who are responsible should be very, very much aware of the need to compensate others.”
>
> **Ghanaian President Nana Akufo-Addo**, to The Associated Press, on the need for the wealthy countries that disproportionately emit greenhouse gases to help less developed countries pay for infrastructure damage and economic losses from climate change. The ongoing climate summit in Egypt is the first to include the idea as a formal agenda item.

wards the past few weeks. On 1 November, Pfizer announced trial results showing its maternal RSV vaccine—which generates antibodies that are passed on to newborns—had nearly 82% efficacy against severe infection in the first 3 months of the baby's life and almost 70% over the first 6 months. The company will now seek regulatory approval. Three days later, the European Commission greenlighted an RSV-targeting antibody called nirsevimab, developed by Sanofi and AstraZeneca, that can be directly injected into newborns. A similar antibody is already approved for infants at high risk of severe disease from RSV because of other conditions, but a phase 3 trial published this year suggests nirsevimab might help all babies. (See a *Science* interview about the discoveries at https://scim.ag/RSVvax.)

NASA postpones Venus mission

PLANETARY SCIENCE | NASA said last week it will delay the launch of its Veritas mission to Venus by 3 years, to 2031, to allow the agency to improve project management after its Psyche spacecraft failed to meet its launch window this summer. Psyche, which will now launch in October 2023 to explore an unusual metallic asteroid, suffered from delays in the development and testing of its flight software at the Jet Propulsion Laboratory (JPL), NASA's lead center for robotic exploration of the Solar System. An independent review commissioned by NASA found JPL has broader problems, such as technicians stretched across too many projects. Postponing Veritas, an orbiter that will map Venus's surface in fine detail, will free up the staff needed to complete Psyche and two multibillion-dollar projects, the Europa Clipper and Mars sample return missions.

Energy agency divides windfall

FUNDING | ITER, the international fusion test reactor under construction in southern France, is the biggest winner from an infusion of extra cash for the U.S. Department of Energy's research, the agency's Office of Science announced last week. ITER and 51 other projects already underway will split $1.55 billion from the Inflation Reduction Act, which President Joe Biden signed in August. The $256 million for ITER will be used to construct parts. Among the office's 10 national laboratories, Oak Ridge is the biggest winner, receiving 32% of the $1.55 billion. The office's total budget is $7.45 billion.

CRISPR therapy subject dies

CLINICAL RESEARCH | A 27-year-old man has died “while participating” in a novel gene-editing trial for his Duchenne muscular dystrophy, according to a foundation his family created to support the experimental treatment. The foundation, Cure Rare Disease, noted the death on 14 October but did not explicitly confirm that the patient, Terry Horgan, received the treatment. In a 1 November statement first reported by The Associated Press, it said its investigation of the cause of death would involve “multiple teams” and could take months. That has led many scientists to believe Horgan had indeed received a planned infusion of viruses carrying DNA that coded for a component of the gene editor CRISPR. The component was an enzyme called Cas9 that normally slices DNA but had been modified to switch on Horgan's gene for dystrophin, a muscle protein. High doses of the same type of modified virus have previously been linked to severe immune reactions and one death in trials of a different therapy for Duchenne, in which the virus is used to carry a replacement gene. Horgan died about 6 weeks after the trial began at the University of Massachusetts Chan Medical School, according to ClinicalTrials.gov.



Most species of harlequin frogs remain critically endangered.

PHOTO: MORLEY READ

CONSERVATION BIOLOGY

Frogs deemed extinct due to fungus live on

A new analysis reveals that over the past 2 decades, scientists have rediscovered one-third of the 87 species of harlequin frogs thought to have disappeared from South and Central America since the 1950s because of a lethal fungus. The work supports hopes that extinctions of the tiny, colorful frogs in the *Atelopus* genus, which comprises 105 known species, were not as widespread as feared. But some of the rediscoveries are based on a single sighting, and collectively the various species may be far from fully recovered, the researchers report this week in *Biological Conservation*. During the past 50 years, the chytrid fungus *Batrachochytrium dendrobatidis* has caused population declines in hundreds of amphibian species, researchers estimate, and harlequin frogs have been among the most severely affected. Understanding why the rediscovered ones survived may help inform conservation efforts. One take-home lesson, according to other researchers: Protecting habitats for all missing amphibians may help them make a comeback.



GLOBAL HEALTH

Invasive mosquito adds to Africa's malaria toll

Anopheles stephensi may dramatically increase the number of people at risk

By **Gretchen Vogel**

Nearly a decade ago, the Republic of Djibouti seemed on the cusp of eliminating malaria. The small country in the Horn of Africa saw only 27 cases in 2012. But between February and May 2013, the disease surged, with 1228 cases, followed in November 2013 by another wave of more than 2100 cases. Strangely, many people fell ill in the capital, Djibouti City. In Africa, malaria is mostly a rural problem.

Far from being an anomaly, the outbreak marked the arrival of a new threat: an invasive mosquito, *Anopheles stephensi*, that appears to have reached Africa from Asia not long before the outbreak. Djibouti saw more than 73,000 malaria cases last year, due at least partly to the mosquito. Last week, a study presented at the annual meeting of the American Society of Tropical Medicine and Hygiene (ASTMH) in Seattle linked an unusual malaria outbreak earlier this year in an Ethiopian city to the same culprit. *An. stephensi* has also been spotted in Sudan, Somalia, and Nigeria, and it may lurk in other countries as well.

"This is one of the biggest movements of a malaria vector that has taken place in the past 50 years," says Seth Irish, a medical entomologist at the World Health Organization (WHO). The evidence is growing that it is adding to the burden of malaria, which already kills half a million people in Africa each year, most of them children under age 5. To better gauge the threat, WHO launched a new initiative in September to step up surveillance for the species and study its habits.

A 2014 paper in *Acta Tropica* about the Djibouti outbreak first reported that *An. stephensi*, native to South Asia and the Arabian Peninsula, had made the short jump to Africa. The paper also raised the alarm about what this might portend: Unlike most of the African mosquitoes that transmit the malaria parasites, *An. stephensi* is a city dweller, which could explain the urban cases—and made the mosquito's arrival a "significant future health threat" for Africa, the researchers warned.

An. stephensi is well-known across its original habitat as an efficient malaria vector, especially in cities. It can transmit both parasites that cause most human malaria cases, *Plasmodium vivax* and the deadlier *P. falciparum*. It thrives in artificial water sources such as cisterns and even deep wells, enabling it to stay active during dry seasons. Its African cousins, such as *An. gambiae* and *An. funestus*, tend to prefer rural environments and lay their eggs in puddles that in many countries only occur in the rainy season, resulting in a respite from the disease during the rest of the year.

***Anopheles stephensi*, a native of South Asia and the Arabian Peninsula, is spreading rapidly in Africa.**

Researchers are still trying to gauge the threat *An. stephensi* poses in Africa. A 2020 study estimated that if the mosquito spreads unchecked, 126 million people could be at increased risk of malaria. However, "We don't know very much about where it is and what the real contribution to transmission is," says Jan Kolaczinski, a medical entomologist at WHO. Epidemiologist Anne Wilson, who has been tracking the spread of *An. stephensi* in Sudan and Ethiopia with her colleagues at the Liverpool School of Tropical Medicine (LSTM), notes that in some areas where *An. stephensi* has been found, malaria cases have not increased.

To better understand the mosquito's role, molecular biologist Fitsum Tadesse of the Armauer Hansen Research Institute in Addis Ababa, Ethiopia, and colleagues tracked cases of malaria in Dire Dawa, a city in the eastern part of the country that had an unusual outbreak of more than 2400 cases during the first half of this year, in the dry season. (The city only had 205 cases in all of 2019.) They tested household members of 80 malaria patients and compared them with households of 210 people who did not have the disease. People living with malaria patients were 5.6 times more

PHOTO: SINCLAIR STAMMERS/SCIENCE SOURCE

likely to be infected, suggesting a nearby source of the disease. Infected households also had more mosquito breeding habitats within 100 meters of their homes, the team found—and 97% of the adult mosquitoes were *An. stephensi*, the team reported at the ASTMH meeting.

The study provides the most direct evidence yet that the invasive insect can cause an increase in malaria cases, says Martin Donnelly, an evolutionary geneticist at LSTM who was not involved in the study. "It is a big step forward," he says.

At the ASTMH meeting, data from entomologist Hmooda Kafy of the University of Khartoum showed that *An. stephensi* occurred in 39 of 61 sites surveyed across the country. In some areas, 88% of households had the mosquito in or near their homes. Kenya, which has stepped up surveillance, has not picked up the species yet but researchers are checking archived samples, Solomon Karoki of the Kenyan Ministry of Health said. Tadesse suspects the mosquito has spread farther than researchers realize, possibly hitching rides in shipping containers: "It's likely you could find it in all corners of the continent," he says.

Although *An. stephensi* is well-adapted to city life, it also breeds in rural cisterns or wells, notes Sarah Zohdy, an entomologist at the U.S. Centers for Disease Control and Prevention and the U.S. President's Malaria Initiative. "We call it an urban vector, but it's really an everywhere vector," she says.

The *An. stephensi* strains found in Africa are largely resistant to the most widely used insecticides, and they seem to prefer to rest in barns or sheds rather than human homes, biting people when they are outside. That means standard mosquito control measures such as insecticide-treated bed nets and indoor insecticide spraying might not be very effective.

One control tactic is to keep water reservoirs covered so that adult mosquitoes can't lay their eggs in them, but this often proves hard to keep up. Another is to add an insecticide to the water that targets the immature mosquitoes in their larval stage. Both approaches also help control *Aedes* mosquitoes, which transmit viral diseases such as dengue and chikungunya. Ethiopia has built its action plan in part on its dengue strategies because of that overlap, says Achamyelesh Mekuanint, a malaria expert at Ethiopia's Ministry of Health.

Irish says more research on *An. stephensi* is urgently needed in order for WHO to fine-tune its messaging and make recommendations for control. "Getting the balance right is important," he says. "It's a huge concern, but we do need to put time and effort into understanding its real impact." ■

ILLUSTRATION: DRUG NARODA/SHUTTERSTOCK AND ANYAPL/SHUTTERSTOCK, ADAPTED BY C. SMITH/*SCIENCE*

SCIENTIFIC COMMUNITY

As Musk reshapes Twitter, academics ponder taking flight

Many researchers are setting up profiles on another social media service known as Mastodon

By **Kai Kupferschmidt**

Mark McCaughrean has been moving his online home in steps. McCaughrean, who is an astronomer at the European Space Agency, has had a profile on Twitter for many years. In the spring, when Elon Musk first suggested buying the social media platform used by nearly 240 million worldwide, many were concerned. Musk calls himself a "free speech absolutist" and promised to stop censoring accounts, raising fears that Twitter would grow nastier and misinformation would drown out reasonable discourse. But for McCaughrean, it was beyond that. "At some level, I made a choice that I don't want to support, personally, his ecosystem."



So McCaughrean decided to open a profile on Mastodon, a recent, much smaller Twitter rival. "I just left a username there," he says. But 2 weeks ago, after the sale went through, McCaughrean started to use the new platform. "I have been much more active there than I have been on Twitter."

With 16,000 followers, McCaughrean is no Twitter celebrity, but he is one of countless scientists who have used the platform to connect with—and debate—colleagues in the same field, as well as scientists from other fields, artists, journalists, and the general public.

Originally dismissed by many as a platform for self-promotion, Twitter has, in recent years, also provided a venue for hate speech, including abuse directed at scientists. But over time, Twitter has also become a major public good, says Michael Bang Petersen, a political scientist at Aarhus University (@M_B_Petersen, 33,000 followers). "I believe it has played important roles in the dissemination of knowledge globally and between scientists and the public during, for example, the pandemic."

Still, with uncertainty about how Twitter will change under Musk, many of the thousands of medical and scientific experts on the platform have started to look for alternatives or are considering giving up on social media altogether. For a while the hashtags #GoodbyeTwitter and #TwitterMigration were trending, and many researchers have been posting their new Mastodon handles, encouraging others to follow them to the site, which has gained more than 500,000 new users within days of Musk completing his purchase.

For the moment, most researchers are waiting to see what happens with Twitter. "I'm hedging my bets with a Mastodon account but not planning to leave in the short term," says biologist Carl Bergstrom (@CT_Bergstrom, 163,000 followers) of the University of Washington, Seattle. Many other researchers are doing the same. That means even if little changes for now, the groundwork is being laid for what could quickly become a digital mass migration of scientists.

The greatest fear is that under Musk discourse on Twitter will deteriorate further. Indeed, as part of massive layoffs at Twitter last week to cut costs, he let go of its curation team, which is largely responsible for quelling misinformation on the platform. This, combined with an exodus of experts, would mean misinformation could go further unchecked. "I have always felt that having expert voices to counter the rampant misinformation is important and necessary," says Boghuma Titanji (@Boghuma),

likely to be infected, suggesting a nearby source of the disease. Infected households also had more mosquito breeding habitats within 100 meters of their homes, the team found—and 97% of the adult mosquitoes were *An. stephensi*, the team reported at the ASTMH meeting.

The study provides the most direct evidence yet that the invasive insect can cause an increase in malaria cases, says Martin Donnelly, an evolutionary geneticist at LSTM who was not involved in the study. "It is a big step forward," he says.

At the ASTMH meeting, data from entomologist Hmooda Kafy of the University of Khartoum showed that *An. stephensi* occurred in 39 of 61 sites surveyed across the country. In some areas, 88% of households had the mosquito in or near their homes. Kenya, which has stepped up surveillance, has not picked up the species yet but researchers are checking archived samples, Solomon Karoki of the Kenyan Ministry of Health said. Tadesse suspects the mosquito has spread farther than researchers realize, possibly hitching rides in shipping containers: "It's likely you could find it in all corners of the continent," he says.

Although *An. stephensi* is well-adapted to city life, it also breeds in rural cisterns or wells, notes Sarah Zohdy, an entomologist at the U.S. Centers for Disease Control and Prevention and the U.S. President's Malaria Initiative. "We call it an urban vector, but it's really an everywhere vector," she says.

The *An. stephensi* strains found in Africa are largely resistant to the most widely used insecticides, and they seem to prefer to rest in barns or sheds rather than human homes, biting people when they are outside. That means standard mosquito control measures such as insecticide-treated bed nets and indoor insecticide spraying might not be very effective.

One control tactic is to keep water reservoirs covered so that adult mosquitoes can't lay their eggs in them, but this often proves hard to keep up. Another is to add an insecticide to the water that targets the immature mosquitoes in their larval stage. Both approaches also help control *Aedes* mosquitoes, which transmit viral diseases such as dengue and chikungunya. Ethiopia has built its action plan in part on its dengue strategies because of that overlap, says Achamyelesh Mekuanint, a malaria expert at Ethiopia's Ministry of Health.

Irish says more research on *An. stephensi* is urgently needed in order for WHO to fine-tune its messaging and make recommendations for control. "Getting the balance right is important," he says. "It's a huge concern, but we do need to put time and effort into understanding its real impact." ■

ILLUSTRATION: DRUG NARODA/SHUTTERSTOCK AND ANYAPL/SHUTTERSTOCK, ADAPTED BY C. SMITH/*SCIENCE*

SCIENTIFIC COMMUNITY

As Musk reshapes Twitter, academics ponder taking flight

Many researchers are setting up profiles on another social media service known as Mastodon

By **Kai Kupferschmidt**

Mark McCaughrean has been moving his online home in steps. McCaughrean, who is an astronomer at the European Space Agency, has had a profile on Twitter for many years. In the spring, when Elon Musk first suggested buying the social media platform used by nearly 240 million worldwide, many were concerned. Musk calls himself a "free speech absolutist" and promised to stop censoring accounts, raising fears that Twitter would grow nastier and misinformation would drown out reasonable discourse. But for McCaughrean, it was beyond that. "At some level, I made a choice that I don't want to support, personally, his ecosystem."



So McCaughrean decided to open a profile on Mastodon, a recent, much smaller Twitter rival. "I just left a username there," he says. But 2 weeks ago, after the sale went through, McCaughrean started to use the new platform. "I have been much more active there than I have been on Twitter."

With 16,000 followers, McCaughrean is no Twitter celebrity, but he is one of countless scientists who have used the platform to connect with—and debate—colleagues in the same field, as well as scientists from other fields, artists, journalists, and the general public.

Originally dismissed by many as a platform for self-promotion, Twitter has, in recent years, also provided a venue for hate speech, including abuse directed at scientists. But over time, Twitter has also become a major public good, says Michael Bang Petersen, a political scientist at Aarhus University (@M_B_Petersen, 33,000 followers). "I believe it has played important roles in the dissemination of knowledge globally and between scientists and the public during, for example, the pandemic."

Still, with uncertainty about how Twitter will change under Musk, many of the thousands of medical and scientific experts on the platform have started to look for alternatives or are considering giving up on social media altogether. For a while the hashtags #GoodbyeTwitter and #TwitterMigration were trending, and many researchers have been posting their new Mastodon handles, encouraging others to follow them to the site, which has gained more than 500,000 new users within days of Musk completing his purchase.

For the moment, most researchers are waiting to see what happens with Twitter. "I'm hedging my bets with a Mastodon account but not planning to leave in the short term," says biologist Carl Bergstrom (@CT_Bergstrom, 163,000 followers) of the University of Washington, Seattle. Many other researchers are doing the same. That means even if little changes for now, the groundwork is being laid for what could quickly become a digital mass migration of scientists.

The greatest fear is that under Musk discourse on Twitter will deteriorate further. Indeed, as part of massive layoffs at Twitter last week to cut costs, he let go of its curation team, which is largely responsible for quelling misinformation on the platform. This, combined with an exodus of experts, would mean misinformation could go further unchecked. "I have always felt that having expert voices to counter the rampant misinformation is important and necessary," says Boghuma Titanji (@Boghuma),

a virologist at Emory University with more than 22,000 Twitter followers.

Others worry the idea of "free speech" will go too far. "While I agree with the importance of free speech on social media, I also worry whether some of Musk's rhetoric on the issue is taken by some users as a relaxation of the norms governing Twitter interactions," Petersen says. "We know from research that the norms governing a social media group do have an effect on the level of hostility in the group."

Indeed, the use of racial slurs on the platform spiked after Musk took over the platform, even though he has said the rules have not changed. "If it becomes too toxic and abusive, I will leave to preserve my well-being and consider other platforms," Titanji says.

The problem of toxicity on the platform only adds to long-standing worries about Twitter's leaders insufficiently protecting some groups of people, especially women and people of color, from harassment and abuse, says Devi Sridhar, a global health expert at the University of Edinburgh. "They rarely acted on reported tweets and there's always been abuse and threats on the platform." Sridhar (@devisridhar, 323,000 followers) says she will see how things develop before deciding to jump ship.

Angela Rasmussen, a virologist at the University of Saskatchewan (@angie_rasmussen, 411,000 followers), has been on the receiving end of such abuse. But she notes that Twitter helped her find her current job and start some scientific collaborations. "Right now, I still find it a useful platform to follow colleagues and learn as well as to share," she says, adding that she won't leave Twitter as long as the good outweighs the bad. "If the people who like to tell me I'm a stupid/fat/ugly/old/unfuckable/unloveable/compromised/corrupt/conflicted/incompetent bitch get a free pass to say whatever without constraint or moderation, the cost-benefit analysis would change for me," she adds.

> **"Right now, I still find it a useful platform to follow colleagues and learn as well as to share."**
>
> **Angela Rasmussen**, University of Saskatchewan

Many researchers, whose tweets add value to the platform for other users, also bristle at the idea of paying a subscription fee to one of the world's richest individuals. Twitter is also rolling out an optional paid service that includes the blue check mark that signals a verified account and fewer ads. "That will definitely push me out the door," Titanji says. "As a matter of principle, I feel social media users are free content creators for these platforms and accessing them should not come at a financial cost to users."

Some of these challenges may become moot if Twitter simply fails as people leave the platform. And although Twitter may be a public good, it has never been a good business: The company has had revenues between $1 billion and $5 billion in recent years, mostly from advertising, but it only ever turned a profit in 2018 and 2019. Musk's attempts to make the business profitable again may well end up dooming the platform, Bergstrom says. "I do think it's a very real possibility that the whole thing collapses in a matter of months to a few years."

But there is a cost to leaving Twitter, too, says Casey Fiesler (@cfiesler, 24,000 followers), an information researcher at the University of Colorado, Boulder, who has studied the migration of online communities. Perhaps the biggest practical consideration for the many researchers who have built a large following on Twitter is that the decision to move elsewhere means starting from the ground up. "Some people have put a huge amount of effort into building a following on Twitter," Fiesler says. "If I do leave, I'm not sure I'd move to Mastodon immediately or just use this as a reason to do less social media," Rasmussen says.

Even so, online migrations tend to be gradual, Fiesler says. In one of her research projects, a participant described it as akin to "watching a shopping mall go slowly out of business." But the speed at which academics are flocking to Mastodon has surprised her. "Things are changing faster than I thought even a week ago," Fiesler says. McCaughrean agrees. "I'm seeing institutions now joining [Mastodon], observatories, institutes," he says. For now, many people will keep a dual presence, Fiesler says; there are already programs that can automatically post on both platforms. For a mass exodus to happen, "there has to be both a compelling reason to leave, and an immediate viable alternative option," she says.

Even if academic Twitter ends up largely moving to Mastodon, the big question is whether the general public will move there, too, allowing scientists to communicate with a broader audience. "When I tweet, I'm talking to my neighbor and the person in the grocery store and the teenager who is thinking about studying science in college," Fiesler says. "That's the beauty of scientists on social media." ■

WORKFORCE

Scientists on trial after speaking out on harassment

Astrophysicist Christian Ott filed a criminal complaint after job offer withdrawn

By **Jeffrey Mervis**

A high-profile sexual harassment case 7 years ago in California is now reverberating in Europe, with implications for those who speak out against the unsavory academic practice of "passing the harasser."

In February 2018, two astrophysicists at the University of Helsinki, Syksy Räsänen and Till Sawala, spearheaded an open letter from more than 70 Finnish astronomers and astrophysicists broadly condemning harassment and discrimination. An accompanying press release also expressed the group's dismay that Christian Ott, a U.S. astrophysicist who was suspended by and subsequently resigned from the California Institute of Technology (Caltech) after it found he had committed gender-based harassment, was about to start a job at Finland's University of Turku.

Combined with similar protests by other scientists, their actions had the desired effect: Within days Turku rescinded its offer and Ott never went to work at its Tuorla Observatory. But next week, a district court in Finland will decide whether the two researchers went too far.

Spurred by a complaint Ott filed with the police 8 months after losing the Turku job, the Finnish government last year charged Räsänen and Sawala with defaming Ott and spreading information that violated his privacy. The two scientists face a substantial fine and a suspended prison sentence if found guilty. The case is one of several legal battles Ott has waged to clear his name, joining other scientists who have turned to the courts after losing jobs or status because of harassment findings.

In December 2017, days before Ott's resignation from Caltech took effect, the University of Stockholm's Nordic Institute for Theoretical Physics offered him a short-term appointment. But pushback from faculty

a virologist at Emory University with more than 22,000 Twitter followers.

Others worry the idea of "free speech" will go too far. "While I agree with the importance of free speech on social media, I also worry whether some of Musk's rhetoric on the issue is taken by some users as a relaxation of the norms governing Twitter interactions," Petersen says. "We know from research that the norms governing a social media group do have an effect on the level of hostility in the group."

Indeed, the use of racial slurs on the platform spiked after Musk took over the platform, even though he has said the rules have not changed. "If it becomes too toxic and abusive, I will leave to preserve my well-being and consider other platforms," Titanji says.

The problem of toxicity on the platform only adds to long-standing worries about Twitter's leaders insufficiently protecting some groups of people, especially women and people of color, from harassment and abuse, says Devi Sridhar, a global health expert at the University of Edinburgh. "They rarely acted on reported tweets and there's always been abuse and threats on the platform." Sridhar (@devisridhar, 323,000 followers) says she will see how things develop before deciding to jump ship.

Angela Rasmussen, a virologist at the University of Saskatchewan (@angie_rasmussen, 411,000 followers), has been on the receiving end of such abuse. But she notes that Twitter helped her find her current job and start some scientific collaborations. "Right now, I still find it a useful platform to follow colleagues and learn as well as to share," she says, adding that she won't leave Twitter as long as the good outweighs the bad. "If the people who like to tell me I'm a stupid/fat/ugly/old/unfuckable/unloveable/compromised/corrupt/conflicted/incompetent bitch get a free pass to say whatever without constraint or moderation, the cost-benefit analysis would change for me," she adds.

> **"Right now, I still find it a useful platform to follow colleagues and learn as well as to share."**
>
> **Angela Rasmussen**, University of Saskatchewan

Many researchers, whose tweets add value to the platform for other users, also bristle at the idea of paying a subscription fee to one of the world's richest individuals. Twitter is also rolling out an optional paid service that includes the blue check mark that signals a verified account and fewer ads. "That will definitely push me out the door," Titanji says. "As a matter of principle, I feel social media users are free content creators for these platforms and accessing them should not come at a financial cost to users."

Some of these challenges may become moot if Twitter simply fails as people leave the platform. And although Twitter may be a public good, it has never been a good business: The company has had revenues between $1 billion and $5 billion in recent years, mostly from advertising, but it only ever turned a profit in 2018 and 2019. Musk's attempts to make the business profitable again may well end up dooming the platform, Bergstrom says. "I do think it's a very real possibility that the whole thing collapses in a matter of months to a few years."

But there is a cost to leaving Twitter, too, says Casey Fiesler (@cfiesler, 24,000 followers), an information researcher at the University of Colorado, Boulder, who has studied the migration of online communities. Perhaps the biggest practical consideration for the many researchers who have built a large following on Twitter is that the decision to move elsewhere means starting from the ground up. "Some people have put a huge amount of effort into building a following on Twitter," Fiesler says. "If I do leave, I'm not sure I'd move to Mastodon immediately or just use this as a reason to do less social media," Rasmussen says.

Even so, online migrations tend to be gradual, Fiesler says. In one of her research projects, a participant described it as akin to "watching a shopping mall go slowly out of business." But the speed at which academics are flocking to Mastodon has surprised her. "Things are changing faster than I thought even a week ago," Fiesler says. McCaughrean agrees. "I'm seeing institutions now joining [Mastodon], observatories, institutes," he says. For now, many people will keep a dual presence, Fiesler says; there are already programs that can automatically post on both platforms. For a mass exodus to happen, "there has to be both a compelling reason to leave, and an immediate viable alternative option," she says.

Even if academic Twitter ends up largely moving to Mastodon, the big question is whether the general public will move there, too, allowing scientists to communicate with a broader audience. "When I tweet, I'm talking to my neighbor and the person in the grocery store and the teenager who is thinking about studying science in college," Fiesler says. "That's the beauty of scientists on social media." ■

WORKFORCE

Scientists on trial after speaking out on harassment

Astrophysicist Christian Ott filed a criminal complaint after job offer withdrawn

By **Jeffrey Mervis**

A high-profile harassment case 7 years ago in California is now reverberating in Europe, with implications for those who speak out against the unsavory academic practice of "passing the harasser."

In February 2018, two astrophysicists at the University of Helsinki, Syksy Räsänen and Till Sawala, spearheaded an open letter from more than 70 Finnish astronomers and astrophysicists broadly condemning harassment and discrimination. An accompanying press release also expressed the group's dismay that Christian Ott, a U.S. astrophysicist who was suspended by and subsequently resigned from the California Institute of Technology (Caltech) after it found he had committed gender-based harassment, was about to start a job at Finland's University of Turku.

Combined with similar protests by other scientists, their actions had the desired effect: Within days Turku rescinded its offer and Ott never went to work at its Tuorla Observatory. But next week, a district court in Finland will decide whether the two researchers went too far.

Spurred by a complaint Ott filed with the police 8 months after losing the Turku job, the Finnish government last year charged Räsänen and Sawala with defaming Ott and spreading information that violated his privacy. The two scientists face a substantial fine and a suspended prison sentence if found guilty. The case is one of several legal battles Ott has waged to clear his name, joining other scientists who have turned to the courts after losing jobs or status because of harassment findings.

In December 2017, days before Ott's resignation from Caltech took effect, the University of Stockholm's Nordic Institute for Theoretical Physics offered him a short-term appointment. But pushback from faculty

led Stockholm officials to reconsider their decision. They contacted Turku, which on 31 January 2018 offered Ott a 2-year contract to start on 1 March.

News of his imminent hiring prompted Räsänen and Sawala to help draft the public statement and a private letter to senior Turku administrators. After Turku announced it was pulling out of the deal, Ott sued both European universities for breach of contract, demanding $1 million in damages, along with reimbursement for lost salary and other expenses. In May 2019, a Swedish court awarded him the equivalent of $66,000, and in March a court in Turku added the equivalent of $89,000.

After looking into Ott's criminal complaint, Finnish government prosecutors decided not to press charges. But Ott appealed, and prosecutors announced in May 2021 the case would go forward. District Judge Stina Selander heard testimony over the summer and is expected to rule on 17 November.

Ott has long maintained that being labeled a harasser has deprived him from working in his chosen field. Among other professional setbacks, Ott says he was forced to resign from the scientific team for the Laser Interferometer Gravitational-Wave Observatory, whose leadership won a Nobel Prize in 2017.

"The publicity destroyed his life, and Räsänen and Sawala were the ringleaders," says his lawyer, Pontus Lindberg. In his testimony, Ott said he hoped the judge's ruling would be "something that will hurt them but will not make it impossible for them to continue with their research."

Speaking last week to *Science*, Sawala's lawyer, Jussi Sarvikivi, said the prosecutor's position appears to be that "any commentary on the Caltech finding demonstrates an intent to harm" Ott because it inevitably casts Ott in a poor light.

During the trial, the defendants' lawyers argued that their clients were relying on "reliable news sources" of what happened at Caltech and had no reason to question their accuracy. The laws under which Räsänen and Sawala are charged also exempt statements about a public figure or someone engaging in a "public activity," a category that includes science.

Räsänen and Sawala declined to comment pending the judge's decision. But in their testimony, they said they were simply speaking out on an important issue facing their profession. "It is the duty of every member of the scientific community to prevent harassment," Räsänen told the judge. "When a harasser can simply move to another institution," Sawala wrote in the 2018 press release, "it is a slap in the face of individuals who suffer harassment."

Caltech is not a party to the case, but its finding that Ott was guilty of gender-based harassment against two graduate students looms over the proceedings. Caltech has issued just a few brief public statements about the case, but newly disclosed documents provide additional details.

In January 2016, Caltech's president, Thomas Rosenbaum, announced that a faculty member had been suspended without pay for the 2015–16 academic year and required to undergo additional mentorship training. It later acknowledged Ott was the subject. Ott then returned to Caltech's payroll in July 2016 on paid leave. That detail, previously unreported, is contained in letters to him from Fiona Harrison, chair of Caltech's physics, math, and astronomy division, that Ott provided to *Science*. In court filings, Ott reported 2017 income of $204,000 from Caltech.

In May 2017, Harrison wrote to Caltech employees that Ott's progress was being monitored and a decision "about [Ott's] possible return" to the faculty would be made in the fall. But 3 weeks earlier, she provided a federal funding agency with more information, according to a letter provided by Ott. Ott would regain regular faculty status at the start of the 2017–18 academic year, Harrison wrote on 27 April to the National Science Foundation, which was funding some of his research. Ott would "work on his research projects, including interacting with students and postdocs," Harrison wrote. Caltech declined to comment on her letters.

But that wasn't the final chapter. On 1 August 2017, a memo from Rosenbaum noted Ott had "made significant progress … [but] remained a divisive element on campus" and that Ott "has decided to resign, effective 31 December."

The 2018 open letter from the Finnish astronomers makes room for what it calls "the possibility of rehabilitation" for harassers if it's preceded by "acknowledgment of the offense and taking responsibility for the harm caused." Ott says he asked repeatedly to sign onto the letter but was rebuffed.

In their testimony, Räsänen and Sawala said Ott refused to answer when asked whether he acknowledged causing harm. And a 2016 complaint Ott filed with the U.S. government places most of the blame for his downfall on his then-employer.



Christian Ott has turned to the courts after Caltech found him to have committed gender-based harassment.

PHOTO: CALIFORNIA INSTITUTE OF TECHNOLOGY

"Caltech's fear of public outcry and potential litigation … led it to botch the investigation of Dr. Ott's prudent and responsible, although certainly not perfect, interactions with the two graduate students," Ott wrote in a filing with the Department of Education's Office for Civil Rights (OCR), which oversees harassment investigations under Title IX. In addition, Ott wrote, "it discriminated against [Ott] because he was a man and the complainants are women."

That complaint never moved forward, Ott told *Science* last week, although he says OCR officials suggested he contact another federal agency that handles allegations of employment discrimination. (OCR doesn't comment on the status of complaints.) "But I decided against it at the time," Ott says, "because Caltech had promised to reinstate me." ■



Farmers in China transplant seedlings for the seasonal rice harvest, which takes weeks of hard work for every hectare.

AGRICULTURE

Perennial rice could be a 'game changer'

Long-term study in China shows yields hold up and farmers save money and time

By **Erik Stokstad**

Grains that grow year after year without having to be replanted could save money, help the environment, and reduce the need for back-breaking labor. Now, the largest real-world test of such a crop—a perennial rice grown in China—is showing promise. Perennial rice can yield harvests as plentiful as the conventional, annually planted crop while benefiting the soil and saving smallholder farmers considerable labor and expense, researchers have found.

"This is the first robust case study" of perennial rice, says Sieglinde Snapp, a soil and crop scientist at the International Maize and Wheat Improvement Center who was not involved with the work. The results show the crop is "a potential game changer," adds Clemens Grünbühel, an ecological anthropologist at the Australian Centre for International Agricultural Research who studies agriculture and rural development. But whether it will catch on is hard to predict, says Susan McCouch, a rice geneticist at Cornell University, because seasonal replanting still has some advantages over the new crop.

All rice is to some extent perennial, sprouting new stems after harvest. The trouble is that this second growth doesn't yield much grain, which is why farmers plow up the paddies and plant new seedlings. The improved perennial rice, in contrast, grows back vigorously for a second harvest. Researchers developed it by crossing an Asian variety of rice with a wild, perennial relative from Nigeria. Improving the offspring took decades, and in 2018 a variety called Perennial Rice 23 (PR23) became commercially available to Chinese farmers.

But how many times PR23 could be harvested before its yield dropped was unclear, as was the size of any economic and environmental benefits. So Fengyi Hu, a geneticist and agronomist at Yunnan University, and others organized longer experiments. They arranged with farmers in three locations to plant the rice and harvest it twice a year for 5 years.

Over 4 years PR23 averaged 6.8 tons of rice per hectare, slightly more than annual rice, they report this week in *Nature Sustainability*. As hoped, the perennial crop grew back again and again without sacrificing the size of the harvest. In the fifth year, however, the yields of PR23 declined, suggesting it needed to be replanted.

Compared with annual rice, the crop left more nutrients in the soil, which also held water better, an important trait for rice grown in regions that depend on rainfall. By next year, Hu says, the researchers hope to know how much greenhouse gas perennial rice farming emits. Existing paddy-grown rice is a major source of methane, for example, which contributes to global warming.

But is the new rice good for farmers? To find out, the researchers compared the effort involved in cultivating PR23 and the annual varieties. Fuel for plowing, the seedlings themselves, and other costs were basically the same the first year, typically $2600 per hectare. But for each following year the perennial rice cost half as much to manage. Each hectare also took between 68 and 77 fewer days of labor.

The rice is catching on in southern China. Yunnan University has provided seed and training to outreach workers, and the total area planted in 2020 quadrupled to 15,333 hectares last year. (That's still a tiny fraction of China's 27 million hectares of rice.) The government also helped promote the crop, Hu says. This year, PR23 is on a list of 29 varieties recommended to farmers by China's Ministry of Agriculture and Rural Affairs.

The largest beneficiary of the labor savings will likely be women and children, who do most of the transplanting of rice seedlings in many rice-growing countries, says Len Wade, an agricultural ecologist at the University of Queensland, St. Lucia, who helped test the rice variety. Mothers will have more time to "look after the family and get the children to school with breakfast and not exhausted," he says. Farmers could also plant abandoned fields and grow more rice, or they might earn more income in side jobs like construction.

Still, researchers note potential risks. Because PR23 enables farmers to till less, fungi and other pathogens can build up in the fields. Insects can persist in the stubble after harvest, then transmit viruses to the regenerating sprouts in the spring. And without tilling, weeds can flourish. Researchers also note that it's more work to resow the perennial rice when its yield falters, because its larger and deeper roots need to be killed.

The potential benefits—and downsides—of the crop will soon come into sharper focus. The perennial rice is being tried in 17 countries in Asia and Africa. A major target is uplands in Asia, where plowing for conventional rice in small, terraced fields hastens soil erosion.

The creators of PR23 "have a proof of concept," Snapp says. "I hope that there's some momentum building." ■

PHOTO: WANG ZHENG VCG/GETTY IMAGES

ARCHAEOLOGY

Iceman's preservation was not a freak event

Body's survival without lucky accidents of climate suggests more ice mummies await

By **Andrew Curry**

In 1991, hikers in the Alps came across a sensational find: a human body, partly encased in ice, at the top of a mountain pass between Italy and Austria. Police initially assumed the man had died in a mountaineering accident, but within weeks archaeologists were arguing he was actually the victim of a 5100-year-old murder.

They were right: Later dubbed Ötzi after the Ötztal Valley nearby, the man's body is the oldest known "ice mummy" on record. His physical condition, equipment, and violent death—confirmed when scans revealed an arrowhead embedded in his shoulder—have opened a window into life in prehistoric Europe. But Ötzi's preservation may not be as unusual as it seemed, archaeologists argued this week. And that could mean more bodies from the distant past are waiting to emerge as the climate warms and ice melts.

Ötzi "was such a huge surprise when he was found people thought he was a freak event," says Lars Pilø, an archaeologist working for the Oppland County Glacier Archaeological Program in Norway. But many of the original assumptions about how weather, climate, and glacial ice conspired to preserve him were wrong, Pilø and other researchers write in the journal *The Holocene*. "This paper sheds new light on the interpretation of this exceptional archaeological find," says Matthias Huss, a glaciologist at ETH Zürich who was not part of the team.

The first archaeologist on the scene 30 years ago was a researcher at the nearby University of Innsbruck named Konrad Spindler. Stunned by the body's remarkable preservation, he came up with a plausible explanation. Damage to Ötzi's backpack and other equipment led Spindler, who died in 2005, to suggest he was fleeing a conflict and had taken refuge in the mountains late in the year. After dying on a high mountain pass, he was quickly covered by winter snow. A climate shift soon sent temperatures plunging for centuries or longer, preserving the body in an icy glacial "time capsule."

Spindler credited the shallow stone gully where hikers found the iceman with shielding him from the relentless flow of glacial ice just a few meters above. The ice must have remained intact until a warm summer in 1991 melted it away, exposing the mummy. "The general understanding was that Ötzi marked this beginning of a cooler period," Huss says, "as people were sure that [he] must have been within the ice without interruption since his death."

But with the retreat of glaciers and ice patches around the world over the past few decades, other ancient remains have emerged, including bodies, hunting equipment, horse manure, and skis. "No one expected similar sites," says Thomas Reitmaier, an archaeologist at the Archaeological Service of the Canton of Grisons in Switzerland and a co-author of the new study. "Now, we have lots, and we find this one fits quite well with the picture of glacial archaeology we've developed."



The iceman was found in a gully, still partly embedded in ice.

PHOTO: LEOPOLD NEKULA SYGMA/GETTY IMAGES

Many of the lucky accidents thought to have preserved Ötzi never happened, the researchers concluded after re-evaluating some 30 years of research on the site and its famous occupant. For example, recent analyses by other researchers of seeds and leaves on and around the body point to a death in spring rather than fall, perhaps leaving Ötzi's body partially exposed in snow over an Alpine summer.

In the centuries that followed, the authors argue, he was repeatedly bared to the elements. Radiocarbon dates from grass, dung, moss, and other organic material from the bottom of the gully are younger than Ötzi's body, an indication that the site was open to the air. "This idea he was frozen in a time capsule isn't right," Pilø says.

That undercuts the idea that a climate shift or cold period set in 5100 years ago, enclosing the body in ice that stayed intact for millennia. "I've myself made this case in courses with students—which I will need to revise," Huss says.

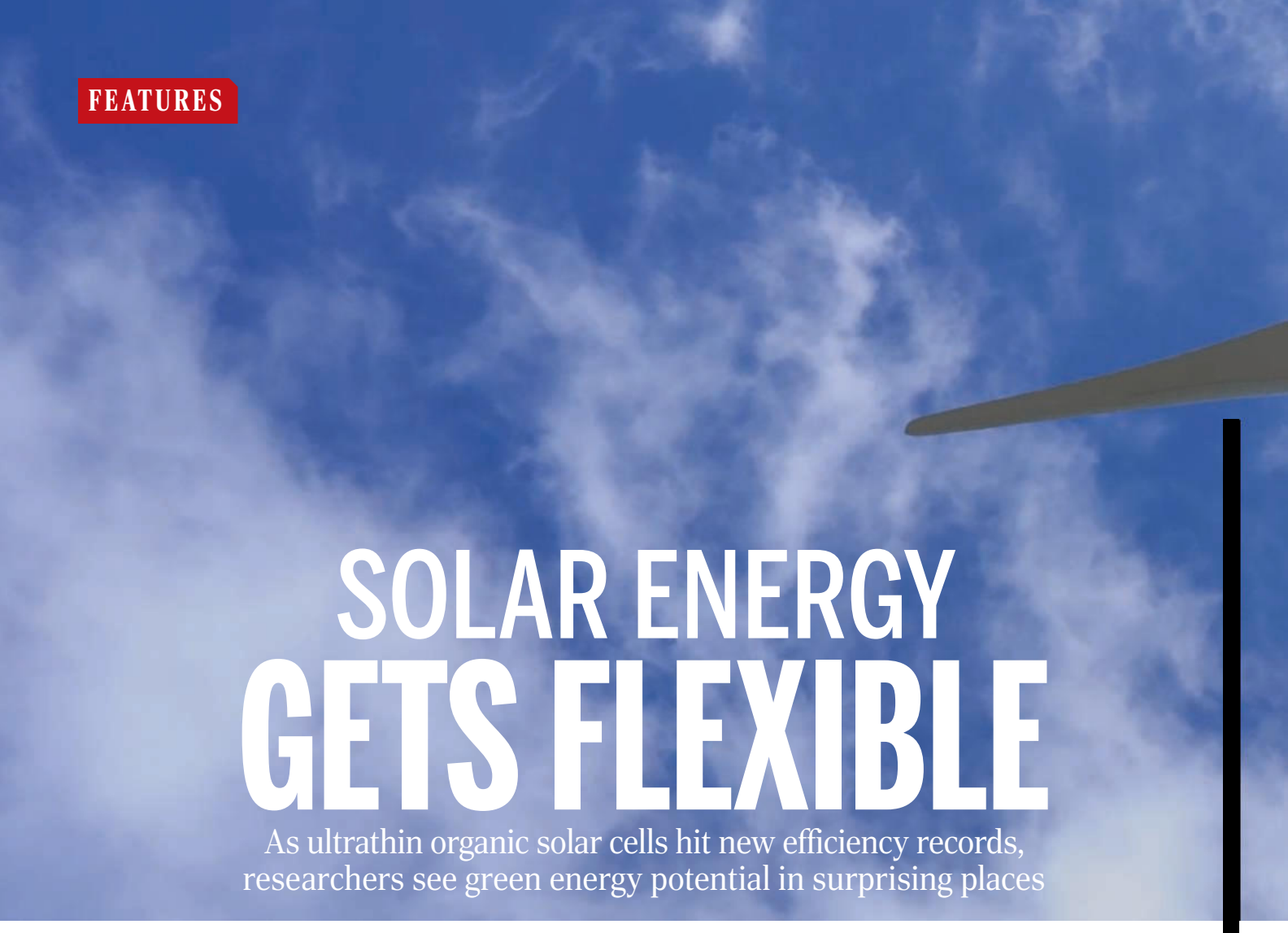
Periodic exposure could also explain why the upper parts of Ötzi's body—particularly the back of his head and his fur cape—are partially decomposed, whereas the lower parts are intact. "If he had been immediately buried in ice he would have been better preserved," Pilø says.

Nor was Ötzi quickly buried where he died, the authors suggest. "There's no way he could have died in the gully," Pilø says. Instead, his missing and scattered belongings, some found 6 meters away, suggest he died on the spring snow above the gully and was later washed into it by meltwater.

That scenario—and not a fight before the mortally wounded victim fled to higher ground—may explain the damaged equipment. Ancient skis, arrowheads, and hunting equipment discovered in Norway, Canada, and elsewhere also show breakage and wear. The similarities suggest shifting ice, or the tumbling of Ötzi's body and equipment by meltwater, splintered or snapped the artifacts.

The new analysis suggests the iceman's dramatic death, shot in the back with an arrow, remains the most unusual aspect of the find. "What is unique—so far—is there was a person shot up there and preserved," Pilø says.

Even that could change. Evidence from other sites in the Alps now suggests mountain passes were often border lines and conflict zones between prehistoric groups. "There's a chance similar sites have preserved parts of human corpses," Reitmaier says. "We have to stay keen in the next years, because ice patches are melting very rapidly everywhere." ■



SOLAR ENERGY GETS FLEXIBLE

As ultrathin organic solar cells hit new efficiency records, researchers see green energy potential in surprising places

By **Robert F. Service**

In November 2021, while the municipal utility in Marburg, Germany, was performing scheduled maintenance on a hot water storage facility, engineers glued 18 solar panels to the outside of the main 10-meter-high cylindrical tank. It's not the typical home for solar panels, most of which are flat, rigid silicon and glass rectangles arrayed on rooftops or in solar parks. The Marburg facility's panels, by contrast, are ultrathin organic films made by Heliatek, a German solar company. In the past few years, Heliatek has mounted its flexible panels on the sides of office towers, the curved roofs of bus stops, and even the cylindrical shaft of an 80-meter-tall windmill. The goal: expanding solar power's reach beyond flat land. "There is a huge market where classical photovoltaics do not work," says Jan Birnstock, Heliatek's chief technical officer.

Organic photovoltaics (OPVs) such as Heliatek's are more than 10 times lighter than silicon panels and in some cases cost just half as much to produce. Some are even transparent, which has architects envisioning solar panels not just on rooftops, but incorporated into building facades, windows, and even indoor spaces. "We want to change every building into an electricity-generating building," Birnstock says.

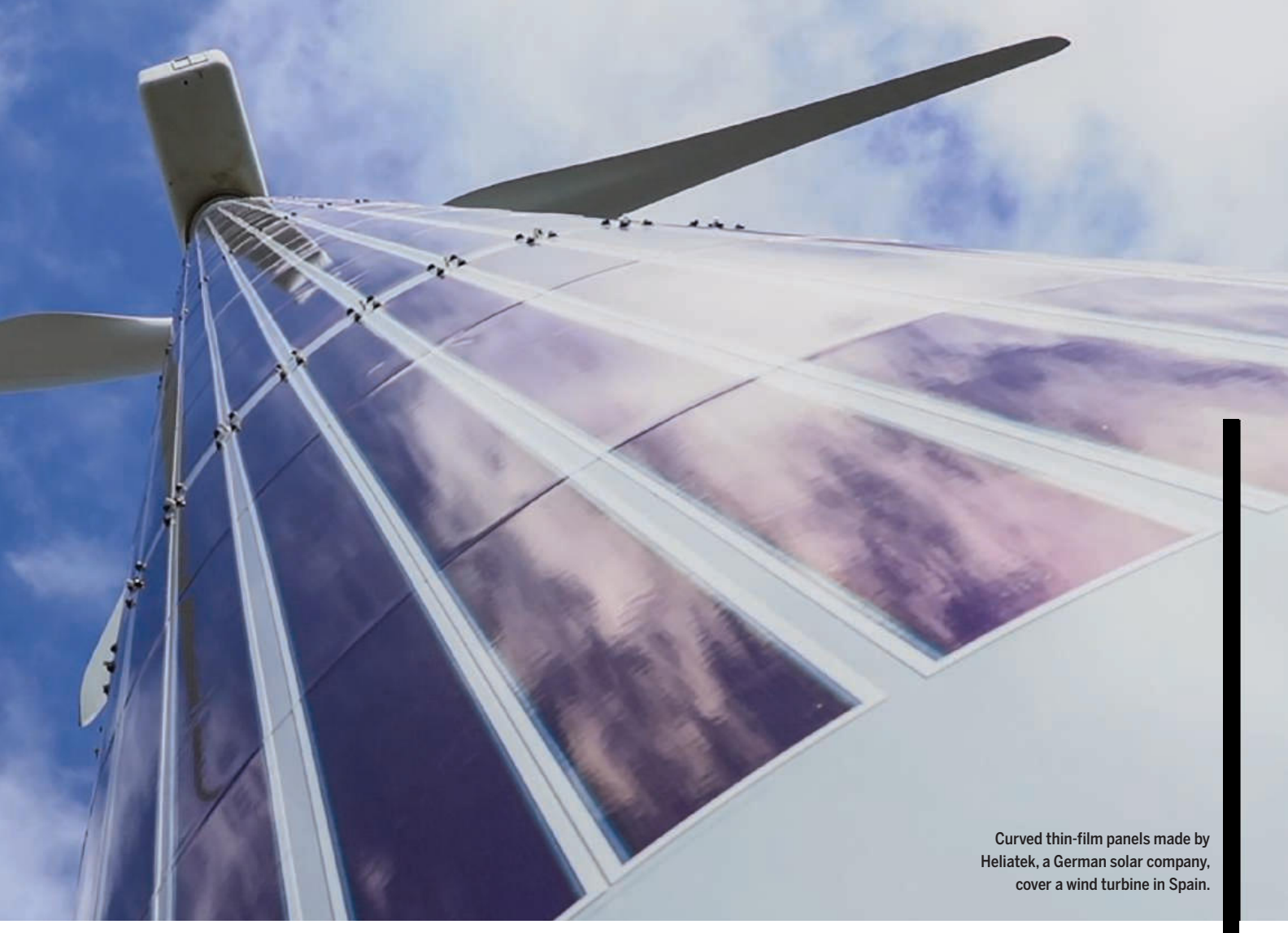
Heliatek's panels are among the few OPVs in practical use, and they convert about 9% of the energy in sunlight to electricity. But in recent years, researchers around the globe have come up with new materials and designs that, in small, labmade prototypes, have reached efficiencies of nearly 20%, approaching silicon and alternative inorganic thin-film solar cells, such as those made from a mix of copper, indium, gallium, and selenium (CIGS). Unlike silicon crystals and CIGS, where researchers are mostly limited to the few chemical options nature gives them, OPVs allow them to tweak bonds, rearrange atoms, and mix in elements from across the periodic table. Those changes represent knobs chemists can adjust to improve their materials' ability to absorb sunlight, conduct charges, and resist degradation. OPVs still fall short on those measures. But, "There is an enormous white space for exploration," says Stephen Forrest, an OPV chemist at the University of Michigan, Ann Arbor.

Even when labmade OPVs look promising, scaling them to create full-size panels remains a challenge, but the potential is enormous. "Right now is a really exciting time in OPVs because the field has made huge leaps in performance, stability, and cost," says Bryon Larson, an OPV expert at the National Renewable Energy Laboratory.

CONVENTIONAL SOLAR POWER—mostly based on silicon—is already a green energy success, supplying roughly 3% of all electricity on the planet. It's the biggest new source of power being added to the grid, with more than 200 gigawatts coming online annually, enough to power 150 million homes. Backed by decades of engineering improvements and a global supply chain, its price continues to drop.

But solar and other green energy sources aren't growing nearly fast enough to meet growing demand and forestall catastrophic climate change. Between the march of

PHOTO: HELIATEK



Curved thin-film panels made by Heliatek, a German solar company, cover a wind turbine in Spain.

global economic development, population growth, and the expected shift of much of the world's cars and trucks from petroleum to electricity, the world's electricity demand is expected to double by 2050. According to the latest estimates from the International Energy Agency, to achieve global net zero carbon emissions by 2050, countries must install renewables at four times the current pace, a challenge the agency calls "formidable." The world needs new sources of renewable power, and fast.

OPV advocates don't see the technology replacing conventional silicon panels for most uses. Rather, they see it helping usher in a wave of new applications and ultimately putting solar in places silicon panels won't work. The field got its start in 1986 when plastic film experts at the Eastman Kodak Company produced the first OPV, which was only 1% efficient at converting the energy in sunlight to electricity. But by the early 2000s, fiddling with the chemical knobs had pushed OPV efficiencies up to about 5%, enough for several companies to try to commercialize them. Their hope was that printing panels on roll-to-roll machines such as newspaper presses would make devices cheap enough to be useful despite their shortcomings. But poor efficiency and degradation under relentless sunlight doomed the early models. "The excitement was there but it was a little too early," Larson says.

Part of the difficulty in raising OPV efficiencies—then as now—is that they work differently from cells made from inorganic materials, such as silicon. All solar cells are sandwichlike devices, with semiconductors in the middle that absorb photons and convert that energy to electrical charges, which then migrate to metallic electrodes layered above and below. When sunlight strikes silicon cells, the added energy kicks electrons out of their orbits around individual silicon atoms, freeing them to flow through the material. Each excited electron leaves behind an electron vacancy, also known as a "hole," which carries a positive charge. The positive charges flow to a negatively charged electrode (the cathode), whereas the electrons flow to a positively charged electrode (the anode), creating an electric current.

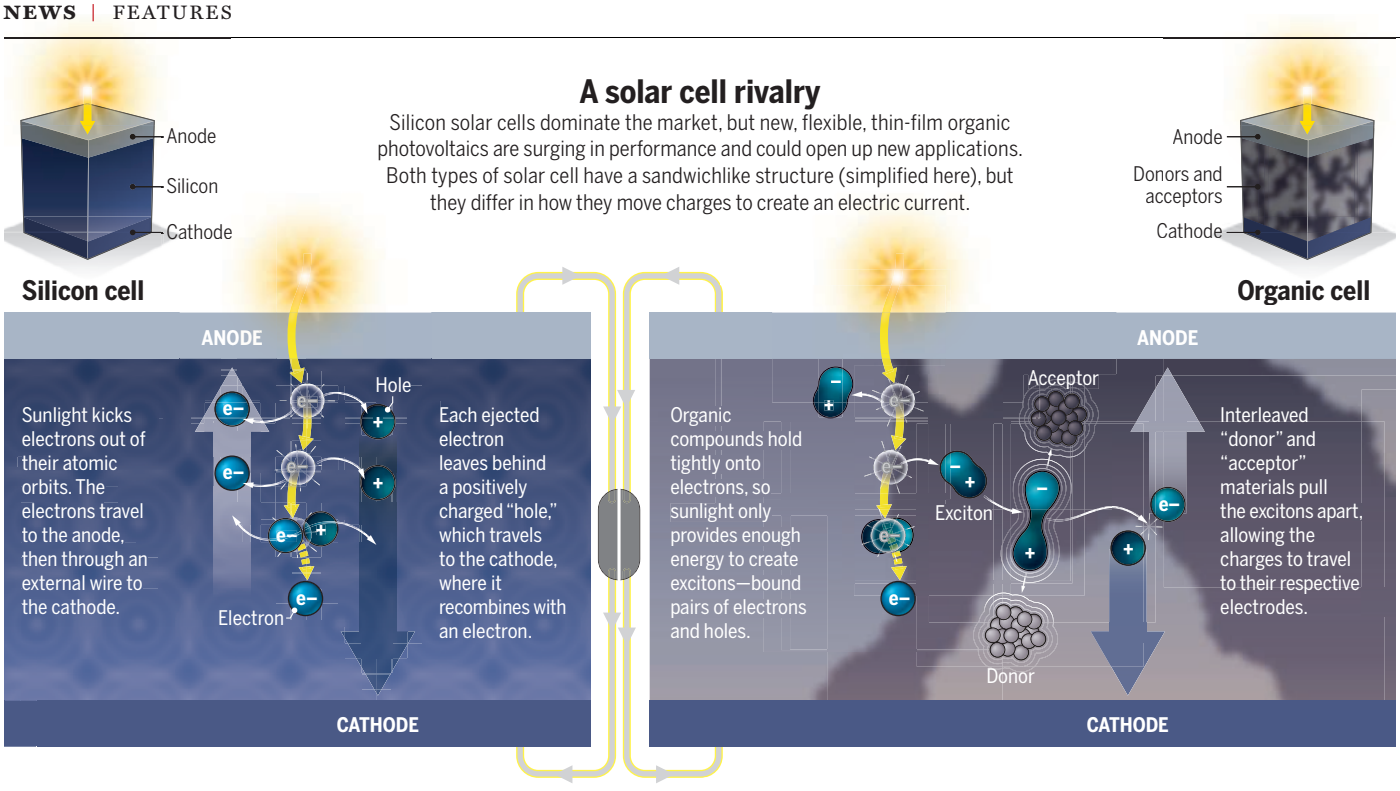
By contrast, the molecules in organic semiconductors tend to hold onto their charges more tightly. When OPVs absorb sunlight, there's enough energy to kick an electron out of its atomic orbit, but not enough for the positive and negative charges to split up and move their separate ways. Rather, these opposite charges stick to each other, creating what is known as an exciton. To generate electricity, the excitons must be separated into positive and negative charges that can travel to their respective electrodes.

The moment of separation comes when excitons move and encounter an interface between two semiconducting components, called donor and acceptor materials. The acceptor attracts electrons, and the donor attracts the positive holes, pulling the exciton apart. It needs to happen quickly: If the excited electron and hole happen to combine with each other before they can reach that interface, they often release their original jolt of excitation as heat, wasting it.

Over the decades, OPV researchers have sought to improve the performance of their devices by coming up with improved

A solar cell rivalry

Silicon solar cells dominate the market, but new, flexible, thin-film organic photovoltaics are surging in performance and could open up new applications. Both types of solar cell have a sandwichlike structure (simplified here), but they differ in how they move charges to create an electric current.



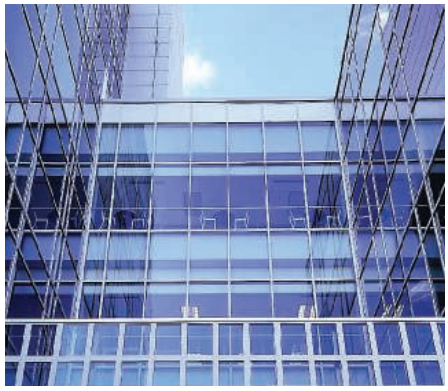
donors and acceptors. Work through the mid-2000s pushed the efficiency above 5%, mainly by incorporating soccer ball–shaped carbon compounds called fullerenes into the materials. The fullerenes' hunger for electrons makes them powerful acceptors. For the next decade, the action shifted to the donors. By 2012, a series of novel semiconducting polymers used as donors propelled efficiencies to 12%.

Then the field suffered a double blow. First, progress plateaued as researchers struggled to find the next breakthrough material. Then a rival thin-film solar technology, called perovskites, burst on the scene. Perovskites are blends of organic and inorganic compounds that are cheap to make, easy to process, and great at capturing sunlight and turning it into electricity. While OPV progress stalled, the efficiency of perovskites skyrocketed from about 6.5% in 2012 to about 24% in 2020. "Perovskites were a stick of dynamite dropped into the OPV world," Larson says. Funding agencies bailed on OPVs and researchers flocked to the hot upstart. "Perovskites were a bandwagon you simply had to be on," says Karl Leo, an OPV researcher at the Technical University of Dresden.

Today, perovskites remain hot. But challenges with long-term stability and their reliance on toxic elements have sapped some enthusiasm. Meanwhile, OPVs soon got a burst of innovation of their own.

In 2015, researchers led by Xiaowei Zhan, a materials scientist at Peking University, reported the first of a new class of nonfullerene acceptors (NFAs). Although fullerenes were good at grabbing and transporting electrons, they were lousy at absorbing sunlight. On a molecular level, Zhan's new compound, dubbed ITIC, looked like an extended Olympic symbol with extra rings, and it did both jobs well, first absorbing red and infrared light and then transporting electrons once excitons split.

Zhan's first NFA device was only about 7% efficient. But chemists around the globe quickly began to tweak ITIC's structure, producing improved versions. By 2016, new NFAs pushed OPV efficiency to 11.5%. By 2018, they hit 16%. And the records keep coming. Last year, Larson and his colleagues reported in *Nature Communications* that by combining multiple donors, an NFA, and a fullerene in a single layer, they created a material that enabled excitons to live longer, and whisked holes more quickly to their electrode, which pushed its efficiency up to 18.4%. And in August, Zhan Lingling at Hangzhou Normal University and her colleagues reported in *Advanced Energy Materials* that an OPV based on a similar multicomponent strategy achieved 19.3% efficiency. "The progress has been really impressive," says Jean-Luc Brédas, an OPV expert at the University of Arizona. "Twenty percent will be reached soon."



Transparent organic photovoltaics are incorporated into the glass facade of the Biomedical and Physical Sciences Building at Michigan State University.

THAT WOULD BRING OPV cells within a few percentage points of their CIGS and silicon rivals. Still, few market watchers believe OPVs will compete head-to-head with silicon anytime soon. Silicon solar cells already command an $85-billion-a-year market, with a 30-year track record and proven durability.

In contrast, OPVs remain niche products. Cheaper OPVs, such as the Heliatek devices, are hampered by low efficiencies, and more efficient ones are still experimental and costly. So, for now, Forrest says, it's best for OPV manufacturers to target new markets where silicon isn't suitable. "If you're competing against silicon, go home, you've already lost," he says.

One fast-growing use is plastering the energy-generating films on the sides of buildings. CIGS and other inorganic thin films can be used the same way. But demand for Heliatek's panels is brisk enough that even though the company only began to sell them last year, it is already building a factory capable of producing 2 million square meters (m^2) annually, enough to provide roughly 200 megawatts of power. Meanwhile, a Swedish company called Epishine sells OPVs that work indoors and can replace

CREDITS: (GRAPHIC) V. ALTOUNIAN/*SCIENCE*; (PHOTO) UBIQUITI

disposable batteries in everything from temperature sensors to automated lighting controls; it has built its own high-speed production line. U.S. startups Ubiquitous Energy and NextEnergy are developing energy-generating OPV windows that primarily capture infrared photons while allowing visible light to pass through, something CIGS and other opaque thin films can't do. And the U.S. Office of Naval Research (ONR) has its eye on using OPVs as power-producing fabrics for tents, backpacks, and other equipment for soldiers on the move. "We want something we can carry to the front," says Paul Armistead, who oversees OPV funding at ONR.

For OPVs to become a significant source of green energy, however, they will need to compete with their rivals on efficiency and durability—and that requires not only new materials, but also manufacturing finesse. The most efficient devices currently exist only as postage stamp–size prototypes in the lab. In theory, scaling up production from 1-square-centimeter cells to 1-m^2 panels is simple. Organics such as polymers and NFAs can be dissolved in solvents and machine-coated over large areas. But each layer in the sandwichlike device must be completely smooth, with few or no imperfections, which can trap moving charges and reduce the overall efficiency. "To get decent efficiencies everything has to work just right," Armistead says.

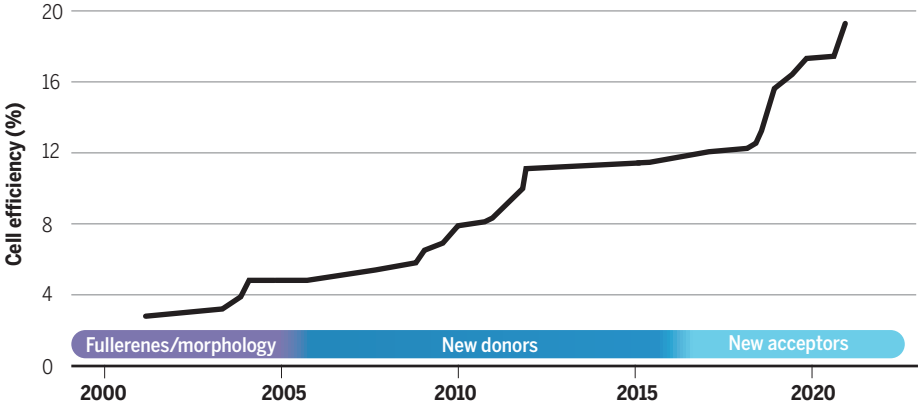
Even more challenging is controlling the makeup of the central layer of the sandwich containing the donors and acceptors. This combination of materials is initially laid down as a liquid with donors, acceptors, sometimes other additives, and solvents all mixed together. As the solvent evaporates, the donors and acceptors segregate, creating two intertwining, continuous networks. The result is a large surface area at the interface between the donor and acceptor regions to separate the charges. The continuous networks also allow the opposite charges to flow along their own paths to the electrodes, with electrons cruising through the network of acceptors and holes moving through the donors.

The intertwining ribbons of donors and acceptors must be extremely thin, because excitons created when photons strike the material can only migrate about 20 nanometers before the charges recombine and the opportunity to generate electricity is lost, says Zhenan Bao, a chemist at Stanford University. "You have to get the morphology right," Armistead says. Doing so reliably, on a large scale, remains a challenge.

He and others are encouraged by a study published on 27 October in *Nature Energy* by Jie Min, an OPV expert at Wuhan University, and his colleagues. Min's team tailored a popular approach for manufacturing thin films at high speed called blade coating. The conventional approach, which mixes donors and acceptors together and spreads the liquid across a moving film and evens it out with what looks like a long squeegee, can produce such films at about 2 m per minute. But by squeegeeing the layers separately one right after the other, the researchers laid down a better network of donors and acceptors at up to 30 m per minute. The resulting cells had efficiencies up to nearly 18%. Min's team also calculates that the faster manufacturing rate could drop OPV costs more than 10-fold and make the price per kilowatt-hour (kWh) competitive with silicon.

Brightening prospects

A 2-decade rise in the efficiency with which organic photovoltaics turn sunlight into electricity was driven at first by molecules called fullerenes and changes to the films' structure, then by better "donor" and "acceptor" materials to separate positive and negative charges.



What remains to be seen, however, is whether such cells will retain the internal structure needed for high efficiency over decades. "In some of the record-breaking cells, the morphology changes over time and the performance doesn't hold up," Armistead says. NFAs are especially susceptible, because the best ones consist of small molecules that can easily shift through the material.

Replacing the NFAs with acceptors woven into long polymers to help keep them in place could help. "They have the chance to be very robust," Armistead says. Progress is on the march here as well. In the 18 August issue of *Advanced Materials*, researchers led by Alex Jen, a materials scientist at the University of Hong Kong, reported all-polymer solar cells that had an efficiency of 17% and retained 90% of their efficiency under accelerated aging tests. "That is quite notable," says Bao, whose team also works on all-polymer cells.

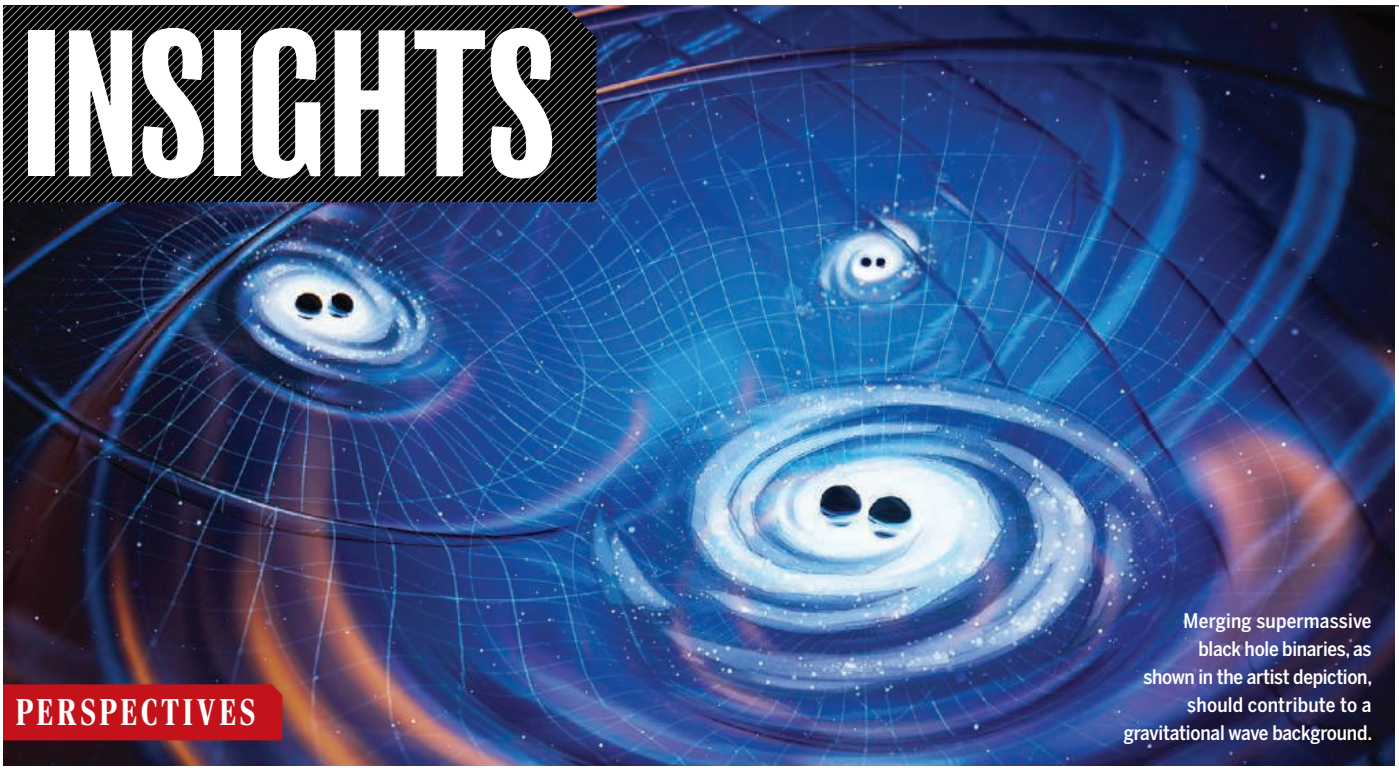
Yet, stability and high efficiency still won't be enough. To make it in the market, solar cells also need to prove reliable for decades. "It's a three-legged stool and you have to have all three legs," Forrest says. Under intense exposure to the ultraviolet (UV) in sunlight, the organics in solar cells can degrade, much as our skin burns during a day at the beach.

In the 14 September 2021 issue of *Nature Communications*, Forrest and his colleagues reported adding a thin layer of UV-absorbing zinc oxide—the same material in some sunscreens—to their OPV, which extended its life up to 30 years in accelerated aging tests. "It's sunscreen for solar cells," Forrest says. Larson, who was not part of Forrest's team, calls it "a huge result."

On one score, OPVs already have a clear advantage over just about every other energy-generating technology: a strikingly low carbon footprint. In evaluating Heliatek's panels, the German testing institute TÜV Rheinland certified that for every kWh of electricity the company's panels produce, at most 15 kilograms (kg) of carbon dioxide (CO_2) would be emitted in making, operating, and eventually disposing of them. That's compared with 49 kg of CO_2/kWh for silicon panels, and a whopping 1008 kg of CO_2/kWh for mining and burning coal. Even with their low efficiencies, Heliatek's panels will generate more than 100 times the energy it takes to make and deal with them over their life span.

OPVs' carbon footprint is sure to lighten further as their efficiency continues to set new records, lifetimes climb, and production methods advance. Those trends are buoying hopes of a world where solar power spreads not only across rooftops and desert scrubland, but also along the curved facades of skyscrapers, the windows of the world, and just about anywhere else people are looking for a bit of juice. That could make prospects for addressing climate change just a little bit brighter. ■

CREDITS: (GRAPHIC) V. ALTOUNIAN/*SCIENCE*; (DATA) ZHANG *ET AL.*, *CHEM. REV.* **18**, 14180 (2022)



Merging supermassive black hole binaries, as shown in the artist depiction, should contribute to a gravitational wave background.

ASTRONOMY

Seeing the gravitational wave universe

Pulsar timing arrays will be a window into the gravitational wave background

By **Chiara M. F. Mingarelli**[1,2] *and* **J. Andrew Casey-Clyde**[1]

Gravitational waves are ripples in the fabric of spacetime that are caused by events such as the merging of black holes. In principle, many types of events occur that could create gravitational waves with frequencies ranging from as high as a few kilohertz to as low as a few nanohertz. Sources of gravitational waves in the nanohertz frequency range include cosmic strings, quantum fluctuations from the early Universe, and, notably, supermassive black hole binaries (SMBHBs). Some gravitational wave sources are so numerous that they are all expected to contribute to a gravitational wave background (GWB). This GWB has been the target of pulsar timing arrays (PTAs) for decades.

PTAs use the correlations between dozens of pulsar pairs to observe the GWB. Recently, the North American Nanohertz Observatory for Gravitational Waves (NANOGrav) (*1*), the European Pulsar Timing Array (EPTA) (*2*), the Parkes Pulsar Timing Array (PPTA) (*3*), and the International Pulsar Timing Array (IPTA) (*4*) have all detected a low-frequency noise in their pulsar data, which may be the first hint of the GWB (see the figure).

The common, low-frequency noise (also called red noise) that the PTAs have measured could be due to the cosmic population of slowly evolving SMBHBs. These SMBHBs create gravitational waves with periods of years to decades in their inspiral phase, the time in the binary's evolution leading to the final merger. This inspiral time scale is very long: A typical equal-mass (1×10^9 solar mass) SMBHB observed with a frequency of 1 nHz is 25 million years from merging. Indeed, these mergers take so long that they should create a stochastic (or random) GWB as a result of the incoherent superposition of potentially tens of thousands of gravitational wave signals. The GWB signal induces delays and advances in the time that it takes for pulses from millisecond pulsars to reach Earth. This signal can be extracted by cross-correlating the residuals—the difference between the expected and the actual arrival time—of pulsar pairs in the PTA. The noise in each pulsar should be independent, whereas the GWB signal should be a common signal in each pulsar—hence, the more pulsar pairs that can be observed, the lower the noise and the larger the signal. The smoking gun of the GWB is the Hellings and Downs curve (*5*), for which we expect the recently detected red noise to eventually conform to a specific functional form (see the figure).

Although strong evidence exists for a common red-noise process (or low-frequency signal) in all the NANOGrav, PPTA, EPTA, and IPTA pulsars, little evidence has been found so far for the Hellings and Downs curve. Whereas Goncharov *et al.* (*6*) concluded after a series of simulations that some common, or similar, red noise originating in pulsars could mimic the common red noise generated by a GWB, Romano *et al.* (*7*) showed that the detection of a common red-noise process should be expected before the Hellings and Downs spatial correlations. If the correlated red noise that is seen in all PTAs truly is a GWB, then detection should be expected with 2 to 5 more years of data (*8*).

Notably, a nanohertz GWB sourced by SMBHBs would indicate that the long-standing final parsec problem—where the SMBHs stall at 1 pc of separation before they efficiently emit gravitational waves—is solved. Having the system stall at a ~1-pc gap

[1]Department of Physics, University of Connecticut, Storrs, CT 06269-3046, USA. [2]Center for Computational Astrophysics, Flatiron Institute, New York, NY 10010, USA. Email: chiara.mingarelli@uconn.edu

IMAGE: OLENA SHMAHALO FOR NANOGRAV

would be almost completely ruled out because the gap depletes the GWB amplitude by ~30% (*9*). Indeed, a GWB amplitude commensurate with the current red noise is so large that it would rule out all but the most optimistic GWB models with no such stalling at 1 pc. For example, Casey-Clyde *et al.* (*10*) found that the number density of SMBHBs in a NANOGrav-like GWB would be five times larger than that in the one predicted by Mingarelli *et al.* (*11*). This either signifies that Mingarelli *et al.* (*11*) were too conservative in their mass and merger rate estimates or that perhaps the merger models need an additional level of sophistication, for example, gas and binary eccentricity, which could in turn increase the number of expected SMBHBs.

Although the focus is on SMBHBs because of their expected presence in the PTA frequency range, other sources are possible. A network of cosmic strings, the existence of which has never been directly demonstrated, is another potential source of a GWB. A third source, a GWB of primordial origin, would provide evidence of an ekpyrotic Universe, where the Big Bang is eventually followed by a Big Crunch. It is not known for sure how long it will take to distinguish between different sources, but Pol *et al.* (*8*) showed that at the time of an initial detection of spatial correlations in pulsar pairs with a signal-to-noise ratio of three, current PTAs should have the capability to distinguish a SMBHB from at least some such exotic sources.

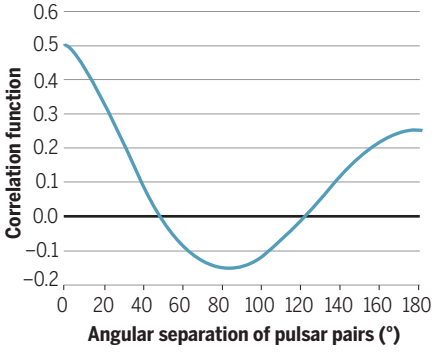
Once the GWB is detected, the next task is to make maps of it, akin to the cosmic microwave background. For instance, individual nearby SMBHB systems and potentially large-scale structures could contribute to or trace the anisotropy in the GWB (*12*). Indeed, GWB anisotropy may enable us to constrain the cosmic population of SMBHBs. Moreover, it will be interesting to see where the anisotropic (excess) power on the sky originates and whether this can be associated with SMBHB activity. However, obtaining upper-limit maps of GWB anisotropy may be challenging because the distribution of pulsars in the sky is itself anisotropic, thwarting the use of the usual spherical harmonics (*13*).

Counterintuitively, detecting continuous gravitational waves from individual inspiraling SMBHB systems by PTAs is possible but more challenging than detecting the GWB. All-sky searches for these continuous waves are computationally expensive and provide poor sky localization for detections. A different path forward is to follow up on binary candidates from electromagnetic surveys, which search for periodic light curves, such as the Catalina Real-time Transient Survey (CRTS). Indeed, recent hydrodynamical simulations predict that periodic light curves could roughly trace the binary's orbit (*14*). Some of these periodic light curves might just be noise that, on short time scales, appears to be periodic. Targeted searches for these binaries appear to be the most promising path forward, because knowing the sky position and rough guess of the binary's period improves PTA sensitivity by an order of magnitude (*15*). As such, extensions to CRTS and the future Rubin Observatory will be crucial for finding possible electromagnetic counterparts to the SMBHB mergers, and facilities such as the next-generation Very Large Array (ngVLA) will be critical for imaging nearby gravitational wave host galaxies.

The detection of the GWB may be imminent, and as such, a new low-frequency era of GW astronomy is at hand. Assuming that the GWB is astrophysical, its detection will likely cast aside any remaining doubt that SMBHs do eventually merge. Moreover, it will yield insights into the expected number density of SMBHBs as a function of redshift, the volume enclosing the GWB, and the minimum mass of a SMBHB that contributes to the background (*10*). All these values are fundamental properties of SMBHBs on which there are extremely limited observational constraints (which also come from PTAs). At present, PTA datasets span about 15 years, and with 5 more years of data, it should be possible to measure a low-frequency turnover in the GWB strain spectrum due to the presence of, for example, gas and stars surrounding the cosmic population of SMBHBs (*8*). Underlying all of this exciting astrophysics will be IPTA datasets formed by combining data from all the major PTAs, substantially increasing detection prospects for all nanohertz gravitational wave sources. ■

Gravitational wave background

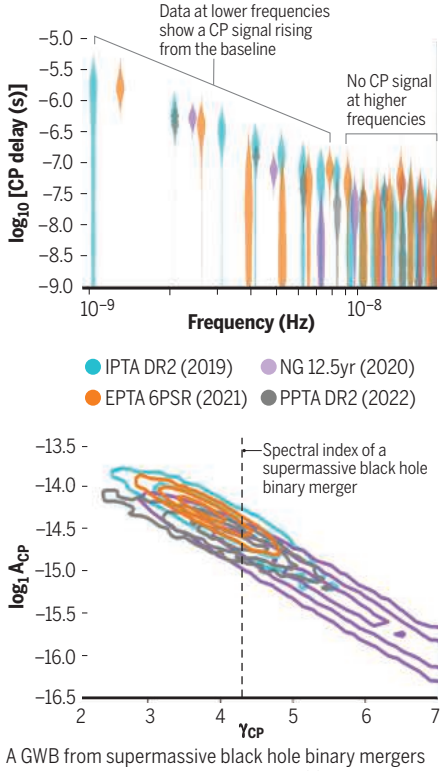
Events such as the merging of supermassive black holes would contribute to a gravitational wave background (GWB) that is potentially detectable by using many pairs of pulsars.



The expected correlation pattern induced by a GWB is called the **Hellings and Downs curve**, which has this specific functional form. Each pulsar pair will appear as a single point on this correlation curve; hence, a credible detection requires a vast number of pulsar pairs. At present, very little evidence of this curve exists in any published pulsar timing array (PTA) dataset.

Common process

Hints of this background show up as low-frequency noise, found in PTAs. Evidence for a common process (CP) red-noise signal in PTA data is highlighted by data release (DR) 2 from the International PTA (IPTA), which incorporates only 9 years of North American Nanohertz Observatory for Gravitational Waves (NANOGrav) data. Combining European PTA (EPTA) and Parkes PTA (PPTA) data is equivalent to getting three additional years of NANOGrav (NG) data.



A GWB from supermassive black hole binary mergers should have a spectral index (γ_{cp}) of 4.33. Identifying both amplitude and spectral index is key to identifying a common process.

REFERENCES AND NOTES

1. Z. Arzoumanian *et al.*, *Astrophys. J. Lett.* **905**, L34 (2020).
2. S. Chen *et al.*, *Mon. Not. R. Astron. Soc.* **508**, 4970 (2021).
3. B. Goncharov *et al.*, *Astrophys. J. Lett.* **917**, L19 (2021).
4. J. Antoniadis *et al.*, *Mon. Not. R. Astron. Soc.* **510**, 4873 (2022).
5. R. W. Hellings, G. S. Downs, *Astrophys. J.* **265**, L39 (1983).
6. B. Goncharov *et al.*, *Astrophys. J. Lett.* **932**, L22 (2022).
7. J. D. Romano, J. S. Hazboun, X. Siemens, A. M. Archibald, *Phys. Rev. D* **103**, 063027 (2021).
8. N. S. Pol *et al.*, *Astrophys. J. Lett.* **911**, L34 (2021).
9. T. Ryu, R. Perna, Z. Haiman, J. P. Ostriker, N. C. Stone, *Mon. Not. R. Astron. Soc.* **473**, 3410 (2018).
10. J. A. Casey-Clyde *et al.*, *Astrophys. J.* **924**, 93 (2022).
11. C. M. F. Mingarelli *et al.*, *Nat. Astron.* **1**, 886 (2017).
12. C. M. F. Mingarelli, T. Sidery, I. Mandel, A. Vecchio, *Phys. Rev. D Part. Fields Gravit. Cosmol.* **88**, 062005 (2013).
13. Y. Ali-Haïmoud, T. L. Smith, C. M. F. Mingarelli, *Phys. Rev. D* **103**, 042009 (2021).
14. B. D. Farris, P. Duffell, A. I. MacFadyen, Z. Haiman, *Astrophys. J.* **783**, 134 (2014).
15. Z. Arzoumanian *et al.*, *Astrophys. J.* **900**, 102 (2020).

ACKNOWLEDGMENTS

We thank P. Baker, S. Chen, J. Lazio, D. Nice, M. McLaughlin, N. Pol, J. Romano, S. Taylor, and S. Vigeland for useful comments. We are supported in part by the National Science Foundation under grant nos. NSF PHY-1748958, PHY-2020265, and AST-2106552. The Flatiron Institute is supported by the Simons Foundation.

10.1126/science.abq1187

GRAPHIC: N. CARY/*SCIENCE* BASED ON (*2*)

FLEXIBLE ELECTRONICS

Connecting liquid metals with sound

A stretchable conductive circuit is formed using a liquid metal-polymer composite

By **Ruirui Qiao**[1] *and* **Shi-Yang Tang**[2]

Gallium (Ga) is a silver-blue metal that melts at 29.8°C and is useful for creating alloys with a wide range of properties. For example, an alloy of Ga and indium (In) with roughly a 3-to-1 ratio has a melting point of just 15.7°C, although In has a melting point of ~157°C. Ga alloys that are liquid at room temperature, called Ga-based liquid metals (GaLMs), generally have low toxicity and high electrical and thermal conductivities (*1*). Because of these properties, GaLM-polymer composite materials have been considered for creating electronic components in stretchable devices, such as those used in biomedical and biosensing applications. However, an approach to reliably fabricate stretchable conductive circuits using GaLM-polymer composites has remained elusive. On page 637 of this issue, Lee *et al.* (*2*) report using acoustic waves to manipulate and create a conductive network of GaLM microdroplets embedded in polymer.

There are two main strategies for making stretchable conductors in flexible electronics—by designing structural patterns to make a nonstretchable material stretchable (akin to cutting holes in a piece of nonstretchable paper to turn it into a stretchable web), and by using materials that are intrinsically stretchable (*3*). The fluidic nature of GaLMs makes them an ideal candidate for the second strategy, because they can be used to build a conductive circuit inside stretchable polymers.

GaLMs react with oxygen to form an oxide skin (1 to 5 nm thick) (*4, 5*), which occurs even at extremely low oxygen concentrations (*6*). This oxide skin is rigid, providing the double benefit of structural stability and the ability for GaLMs to adhere to surfaces through electrostatic interactions (*5*). Because of these properties, a flexible circuit can be patterned onto a stretchable polymer substrate simply by printing GaLMs through a nozzle, spray painting over a stencil, or even stamping with a mold (*1, 7*). However, the oxide skins also pose a set of practical problems for the use of GaLMs. When printing with GaLM droplets suspended in a carrier fluid (*7, 8*), the droplets often do not merge back together, owing to the presence of the oxide skin. As a result, structures formed by these droplets with the oxide skin tend to have poor conductivity and require breaking the oxide skin and recombining the droplets together to form conductive pathways (*8*).

To eliminate the need to recombine the droplets, Lee *et al.* used an acoustic field to help connect individual GaLM microdroplets embedded in polymer (see the figure). With the application of acoustic waves at an ultrasound frequency of 20 kHz, the microdroplets reflect the acoustic energy and shed off nanosized droplets that bridge the microdroplets together. It was surprising that the chain of GaLM microdroplets and nanodroplets form nearly perfect electrical contacts—without merging with the microdroplets or with each other, and without removing the oxide skins. When the composite is stretched, the microdroplet-nanodroplet chain remains intact as the microdroplets change their shape to accommodate the stretch.

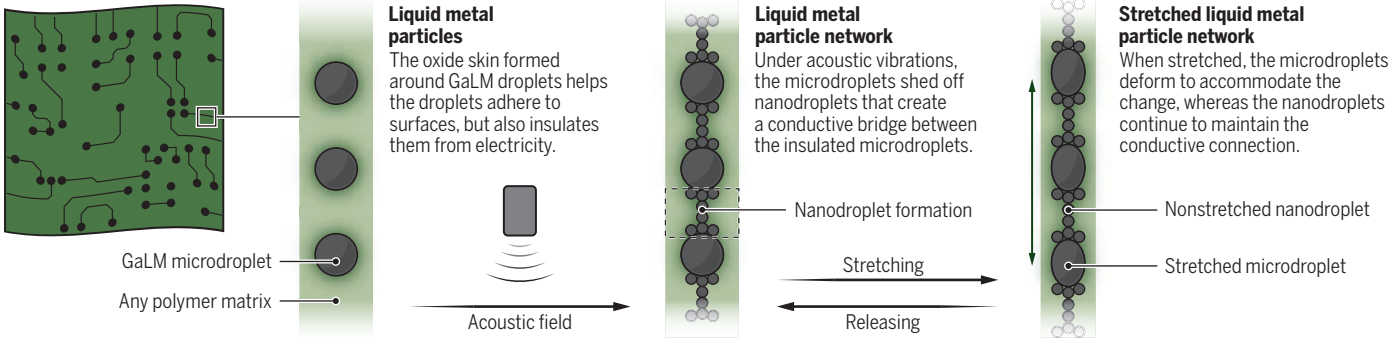
Previous creations of GaLM-polymer composites also face leakage issues under mechanical stress. The leaked GaLMs are difficult to clean and may diffuse into other metallic components inside the device, critically damaging its circuits. To address this problem, Lee *et al.* used smaller GaLM microdroplets (~2 μm), which decreases the chance of breaking and spilling the embedded droplets. The toughness of the composite is enhanced by the GaLM chains, which help absorb mechanical stress. Lee *et al.* demonstrate the versatility of their method for forming conductors using different polymer matrices and other types of liquid metals.

Despite the many special properties of GaLMs and the attention they have received in recent years within the soft electronics field (*9, 10*), GaLMs have yet to be widely implemented by the electronics industry. GaLMs remain a relatively expensive material, with a price tag of ~$0.50/g in 2022. This is ~200 times the price of aluminum and ~66 times the price of copper, which are common conductors in soft electronics. In addition, the electrical conductivity of GaLMs is only ~1/10 that of aluminum and ~1/17 that of copper. Thus, using GaLMs for making simple electronic components is not cost-effective. There may be ways to improve interactions

[1]Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, Brisbane, QLD, Australia. [2]Department of Electronic, Electrical and Systems Engineering, University of Birmingham, Birmingham, UK. Email: r.qiao@uq.edu.au; s.tang@bham.ac.uk

A liquid metal network inside a polymer matrix

Gallium-based liquid metals (GaLMs) are useful for creating complex circuits inside stretchable polymers because of their low melting point, which allows them to be printed using a nozzle, although several practical challenges remain.



GRAPHIC: A. FISHER/SCIENCE

between GaLMs and polymer matrices by using chemistry strategies (*11*), such as selecting polymers that can bind to liquid metal surfaces. Additionally, the suppression of surface oxidation by chemicals (e.g., phosphoric acid) may also enhance the conductivity of GaLMs (*11*). For specific applications that require flexibility, conductors made from GaLM composites have the advantage of being able to maintain conductivity when stretched to more than three times their length (*12*, *13*). This is often regarded as the most important property that GaLM composites can offer.

More research of the interfacial properties, including those between the liquid metal core and the oxide skin, and between the GaLM and the polymer, is essential to further the technology. The method presented by Lee *et al.* helps to overcome a major challenge in creating conductive circuits with GaLM-polymer composites, but the composites still face a number of manufacturing challenges. For example, processes such as sonication and stir-mixing in a carrier fluid for creating GaLM microdroplets tend to lack precise control over the size of the produced droplets. This is a problem because the size of the droplets directly affects the electrical and mechanical performance of the composites.

The biocompatibility and electrical properties of these composites make them particularly attractive for biomedical applications. However, in the daily use of stretchable electronics such as wearable sensors, a scenario that demands the extreme strain parameters offered by GaLM-polymer composites is rare. For reference, the human skin begins to fracture when stretched to ~2.5 times its length (*14*). Other applications such as triboelectric energy harvesters that can generate electricity from being stretched and released may be possible by tuning the electromechanical properties of GaLM composites. Such applications and other yet-to-be-explored avenues may more effectively utilize the full range of the stretchability of the composites. ■

REFERENCES AND NOTES

1. S.-Y. Tang *et al.*, *Annu. Rev. Mater. Res.* **51**, 381 (2021).
2. W. Lee *et al.*, *Science* **378**, 637 (2022).
3. N. Matsuhisa *et al.*, *Chem. Soc. Rev.* **48**, 2946 (2019).
4. T. Daeneke *et al.*, *Chem. Soc. Rev.* **47**, 4073 (2018).
5. M. D. Dickey, *ACS Appl. Mater. Interfaces* **6**, 18369 (2014).
6. T. Liu *et al.*, *J. Microelectromech. Syst.* **21**, 443 (2011).
7. S. Chen, J. Liu, *iScience* **24**, 102026 (2021).
8. N. Kazem *et al.*, *Adv. Mater.* **29**, 1605985 (2017).
9. Z. Ma *et al.*, *Nat. Mater.* **20**, 859 (2021).
10. P. Won *et al.*, *iScience* **24**, 102698 (2021).
11. S.-Y. Tang, R. Qiao, *Accounts Mater. Res.* **2**, 966 (2021).
12. M. J. Ford, *et al.*, *Adv. Mater.* **32**, 2002929 (2020).
13. J. E. Park *et al.*, *Adv. Mater.* **32**, 2002178 (2020).
14. H. Joodaki, M. B. Panzer, *Proc. Inst. Mech. Eng. H* **232**, 323 (2018).

ACKNOWLEDGMENTS

R.Q. receives funding support from the National Health and Medical Research Council (grant no. APP1196850).

10.1126/science.ade1813

ENERGY POLICY

Toward a low-carbon transition in India

Electricity sector policies should be designed not only to mitigate carbon emissions but also to reduce inequities

By **Ranjit Deshmukh**[1] *and* **Sushanta Chatterjee**[2]

In 2021, 40% of India's global greenhouse gas emissions came from electricity generation, mainly powered by coal plants (*1*). Air pollution from these power plants, composed of predominantly sulfur dioxide and particulate matter, is one of the leading causes of respiratory and heart diseases in India, which result in ~80,000 premature deaths annually (*2*). Understanding the impacts of electricity sector policies aimed at mitigating carbon emissions and air pollution is critical for addressing these climate and public health crises. On page 618 of this issue, Sengupta *et al.* (*3*) examine how carbon taxes, balancing electricity generation and consumption across larger regions, and sulfur-control regulations affect near-term costs and emissions of India's electricity sector. They find that these policies can cause inequalities in air pollution exposure across different regions.

A tax levied against power plants based on their carbon emissions—i.e., a carbon tax—may be an effective policy instrument for controlling carbon emissions in high-income countries (*4*). However, this strategy may not have the same effect on low- and middle-income countries, such as India. In theory, imposing a carbon tax on the electricity sector would make both coal power plants more expensive to operate and alternative lower-carbon energy sources more competitive. However, India has limited natural gas resources or energy storage capacity tied with renewable energy that can adequately substitute for coal electricity generation (*5*). According to Sengupta *et al.*, a carbon tax in India is unlikely to meaningfully reduce coal electricity generation without resulting in blackouts, at least in the near term.

Although India has a well-connected electricity grid, most electricity supply and demand is balanced separately by each state. However, balancing electricity supply and demand beyond the state level can reduce costs by optimally using the lowest-cost power plants. Power plants in the coal-producing eastern states are, on average, less efficient (i.e., they have higher emissions per unit of energy generation) but more cost-effective than those in the rest of the country because of their proximity to coal mines. According to Sengupta *et al.*, expanding the balancing regions of the electric grid to neighboring states will enable the cheaper coal power plants in the eastern states to sell more of their electricity, resulting in a decrease in total costs but also a modest increase in carbon and sulfur emissions. Although this strategy may result in extra emissions in the near term, balancing the supply and demand over larger regions will help manage the variability of weather-dependent solar and wind energy generation and increase their adoption (*6*).

With the rising energy demand in India, the low-cost coal plants in the poorer coal-producing regions are likely to continue operating—even as the rest of India transitions to greener energy sources. This will perpetuate the social inequality that stems from communities in these regions being already poorer and disproportionately exposed to air pollution from coal generation compared with the rest of India. Several strategies could limit sulfur and particulate matter emissions from coal power plants and improve public health in these regions. For instance, installing pollution-control equipment called flue gas desulfurizers (FGDs) can capture sulfur dioxide and particulate matter from coal power plants. However, although the Indian government adopted a mandate requiring coal power plants to install FGDs in 2015, only 5% of facilities had done so by 2021. To better triage this problem, coal power plants causing the greatest public exposure to air pollution should be prioritized for FGD installation. The potential electricity rate increases resulting from FGDs could be spread across all taxpayers or electricity buyers, which will help provide economic relief, especially for the poorer coal-producing states (*7*). This may also help stabilize the revenue stream for coal power plant operators and help them adapt technologies, such as FGDs. Another strategy

[1]Environmental Studies Program and the Bren School of Environmental Science and Management, University of California, Santa Barbara, CA, USA. [2]Central Electricity Regulatory Commission, New Delhi, India. Email: rdeshmukh@ucsb.edu; chatterjee.sushanta@gmail.com

between GaLMs and polymer matrices by using chemistry strategies (*11*), such as selecting polymers that can bind to liquid metal surfaces. Additionally, the suppression of surface oxidation by chemicals (e.g., phosphoric acid) may also enhance the conductivity of GaLMs (*11*). For specific applications that require flexibility, conductors made from GaLM composites have the advantage of being able to maintain conductivity when stretched to more than three times their length (*12*, *13*). This is often regarded as the most important property that GaLM composites can offer.

More research of the interfacial properties, including those between the liquid metal core and the oxide skin, and between the GaLM and the polymer, is essential to further the technology. The method presented by Lee *et al.* helps to overcome a major challenge in creating conductive circuits with GaLM-polymer composites, but the composites still face a number of manufacturing challenges. For example, processes such as sonication and stir-mixing in a carrier fluid for creating GaLM microdroplets tend to lack precise control over the size of the produced droplets. This is a problem because the size of the droplets directly affects the electrical and mechanical performance of the composites.

The biocompatibility and electrical properties of these composites make them particularly attractive for biomedical applications. However, in the daily use of stretchable electronics such as wearable sensors, a scenario that demands the extreme strain parameters offered by GaLM-polymer composites is rare. For reference, the human skin begins to fracture when stretched to ~2.5 times its length (*14*). Other applications such as triboelectric energy harvesters that can generate electricity from being stretched and released may be possible by tuning the electromechanical properties of GaLM composites. Such applications and other yet-to-be-explored avenues may more effectively utilize the full range of the stretchability of the composites. ■

REFERENCES AND NOTES

1. S.-Y. Tang *et al.*, *Annu. Rev. Mater. Res.* **51**, 381 (2021).
2. W. Lee *et al.*, *Science* **378**, 637 (2022).
3. N. Matsuhisa *et al.*, *Chem. Soc. Rev.* **48**, 2946 (2019).
4. T. Daeneke *et al.*, *Chem. Soc. Rev.* **47**, 4073 (2018).
5. M. D. Dickey, *ACS Appl. Mater. Interfaces* **6**, 18369 (2014).
6. T. Liu *et al.*, *J. Microelectromech. Syst.* **21**, 443 (2011).
7. S. Chen, J. Liu, *iScience* **24**, 102026 (2021).
8. N. Kazem *et al.*, *Adv. Mater.* **29**, 1605985 (2017).
9. Z. Ma *et al.*, *Nat. Mater.* **20**, 859 (2021).
10. P. Won *et al.*, *iScience* **24**, 102698 (2021).
11. S.-Y. Tang, R. Qiao, *Accounts Mater. Res.* **2**, 966 (2021).
12. M. J. Ford, *et al.*, *Adv. Mater.* **32**, 2002929 (2020).
13. J. E. Park *et al.*, *Adv. Mater.* **32**, 2002178 (2020).
14. H. Joodaki, M. B. Panzer, *Proc. Inst. Mech. Eng. H* **232**, 323 (2018).

ACKNOWLEDGMENTS

R.Q. receives funding support from the National Health and Medical Research Council (grant no. APP1196850).

10.1126/science.ade1813

ENERGY POLICY

Toward a low-carbon transition in India

Electricity sector policies should be designed not only to mitigate carbon emissions but also to reduce inequities

By **Ranjit Deshmukh**[1] *and* **Sushanta Chatterjee**[2]

In 2021, 40% of India's global greenhouse gas emissions came from electricity generation, mainly powered by coal plants (*1*). Air pollution from these power plants, composed of predominantly sulfur dioxide and particulate matter, is one of the leading causes of respiratory and heart diseases in India, which result in ~80,000 premature deaths annually (*2*). Understanding the impacts of electricity sector policies aimed at mitigating carbon emissions and air pollution is critical for addressing these climate and public health crises. On page 618 of this issue, Sengupta *et al.* (*3*) examine how carbon taxes, balancing electricity generation and consumption across larger regions, and sulfur-control regulations affect near-term costs and emissions of India's electricity sector. They find that these policies can cause inequalities in air pollution exposure across different regions.

A tax levied against power plants based on their carbon emissions—i.e., a carbon tax—may be an effective policy instrument for controlling carbon emissions in high-income countries (*4*). However, this strategy may not have the same effect on low- and middle-income countries, such as India. In theory, imposing a carbon tax on the electricity sector would make both coal power plants more expensive to operate and alternative lower-carbon energy sources more competitive. However, India has limited natural gas resources or energy storage capacity tied with renewable energy that can adequately substitute for coal electricity generation (*5*). According to Sengupta *et al.*, a carbon tax in India is unlikely to meaningfully reduce coal electricity generation without resulting in blackouts, at least in the near term.

Although India has a well-connected electricity grid, most electricity supply and demand is balanced separately by each state. However, balancing electricity supply and demand beyond the state level can reduce costs by optimally using the lowest-cost power plants. Power plants in the coal-producing eastern states are, on average, less efficient (i.e., they have higher emissions per unit of energy generation) but more cost-effective than those in the rest of the country because of their proximity to coal mines. According to Sengupta *et al.*, expanding the balancing regions of the electric grid to neighboring states will enable the cheaper coal power plants in the eastern states to sell more of their electricity, resulting in a decrease in total costs but also a modest increase in carbon and sulfur emissions. Although this strategy may result in extra emissions in the near term, balancing the supply and demand over larger regions will help manage the variability of weather-dependent solar and wind energy generation and increase their adoption (*6*).

With the rising energy demand in India, the low-cost coal plants in the poorer coal-producing regions are likely to continue operating—even as the rest of India transitions to greener energy sources. This will perpetuate the social inequality that stems from communities in these regions being already poorer and disproportionately exposed to air pollution from coal generation compared with the rest of India. Several strategies could limit sulfur and particulate matter emissions from coal power plants and improve public health in these regions. For instance, installing pollution-control equipment called flue gas desulfurizers (FGDs) can capture sulfur dioxide and particulate matter from coal power plants. However, although the Indian government adopted a mandate requiring coal power plants to install FGDs in 2015, only 5% of facilities had done so by 2021. To better triage this problem, coal power plants causing the greatest public exposure to air pollution should be prioritized for FGD installation. The potential electricity rate increases resulting from FGDs could be spread across all taxpayers or electricity buyers, which will help provide economic relief, especially for the poorer coal-producing states (*7*). This may also help stabilize the revenue stream for coal power plant operators and help them adapt technologies, such as FGDs. Another strategy

[1]Environmental Studies Program and the Bren School of Environmental Science and Management, University of California, Santa Barbara, CA, USA. [2]Central Electricity Regulatory Commission, New Delhi, India. Email: rdeshmukh@ucsb.edu; chatterjee.sushanta@gmail.com



Power plant air pollution limits and job creation in coal-producing regions are key to India's clean-energy goals.

is to encourage investment in solar and wind power plants in the coal-producing regions. These energy infrastructure investments will also provide local job opportunities as the regions transition away from fossil fuel–based power generation in the long term (*8*).

India faces the twin challenges of mitigating carbon emissions and meeting an increasing energy demand. Several policies and regulations have been introduced to reduce overall energy demand and increase the supply for zero-carbon electricity. For instance, the government has undertaken large-scale procurement of energy-efficient home and office appliances to decrease their prices. It has also created a market for energy-saving certificates, where businesses saving more energy than their targets can sell the leftover credit to another company, creating a monetary incentive for businesses to meet energy efficiency targets (*9*). In addition, it has introduced competition through bidding by renewable energy companies, targets for each state to purchase renewable energy, exemptions for renewable energy from transmission charges (*10*), and transmission infrastructure for renewable energy (*11*). These policies have since led to some of the world's lowest solar and wind energy prices ($30 to $40 USD per megawatt-hour) (*12*). By 2021, India had become the world's fifth-largest solar power producer, with a capacity of 50 GW, and the fourth-largest wind power producer, with a capacity of 40 GW (*13*).

Although the carbon emissions and gross domestic product per capita of India are still less than half the global average, the country has ambitious plans for mitigating its carbon emissions. At the 26th United Nations Climate Change Conference held in 2021 in Glasgow, UK, the Indian government pledged a net-zero emissions target by 2070 and a near-term target of producing 50% of its electricity from renewable energy by 2030 (*14*). For India, the pursuit of climate and renewable energy targets is important, as is public health, employment, and energy affordability across regions and communities to ensure equitable growth. As India continues to develop its economy, balancing the near-term and long-term effects of its electricity sector policies, as well as their impact on social inequalities, will be critical to ensure a low-carbon transition that is green as well as just. ■

REFERENCES AND NOTES

1. R. R. Mohan *et al.*, "Greenhouse Gas Emission Estimates from the Energy Sector in India at the Sub-national Level (version 2.0)" (GHG Platform India, 2019).
2. M. Cropper *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2017936118 (2021).
3. S. Sengupta *et al.*, *Science* **378**, eabh1484 (2022).
4. J. E. Stiglitz, N. Stern, "Report of the High-Level Commission on Carbon Pricing" (Carbon Pricing Leadership Coalition, 2017); https://bit.ly/3TkiiLP.
5. R. Deshmukh, A. Phadke, D. S. Callaway, *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2008128118 (2021).
6. E. Ela *et al.*, *Electr. J.* **29**, 51 (2016).
7. "Analysis of Factors Impacting Retail Tariff and Measures to Address Them" (Forum of Regulators, 2021); https://bit.ly/3EYieNy.
8. S. Pai, H. Zerriffi, J. Jewell, J. Pathak, *Environ. Res. Lett.* **15**, 034065 (2020).
9. "Roadmap of Sustainable and Holistic Approach to National Energy Efficiency" (Bureau of Energy Efficiency, 2019); https://beeindia.gov.in/sites/default/files/Roshanee_print%20version%282%29.pdf.
10. "Tariff Policy" (Ministry of Power, 2016); https://bit.ly/3eJHX1q.
11. "Report on Green Energy Corridors – II: Part-A" (Power Grid Corporation of India Ltd., 2022); www.powergrid.in/sites/default/files/footer/smartgrid/Green%20Energy%20Corridor%202-Part%20A.pdf.
12. "Renewable Power Generation Costs in 2020" (International Renewable Energy Agency, 2021); https://bit.ly/3Teskhw.
13. "Renewable Capacity Statistics 2022" (International Renewable Energy Agency, 2022); www.irena.org/publications/2022/Apr/Renewable-Capacity-Statistics-2022.
14. Prime Minister's Office, "National Statement by Prime Minister Shri Narendra Modi at COP26 Summit in Glasgow" (Press Information Bureau of India, 2021); https://bit.ly/3goRWcS.

ACKNOWLEDGMENTS

The views of the authors are personal and do not represent the views of the Central Electricity Regulatory Commission.

10.1126/science.ade6040

MARINE CONSERVATION

Good and bad news for ocean predators

Some tunas and billfishes are recovering, but sharks continue to decline

By **Matthew G. Burgess**[1,2,3] *and* **Sarah L. Becker**[1,2]

As human population and economies have grown rapidly over the past 100 years, ecosystems worldwide have faced increasing pressure from overexploitation, habitat destruction, and other threats (*1*). In the oceans, roughly half of all commercially harvested fish and invertebrate stocks became overfished during the 20th century (*2*), and larger predators, such as billfishes and sharks, also dwindled (*3*). The 21st century has seen some marine fish and invertebrate stocks begin recovering owing to management efforts (*4*), whereas poorly managed stocks continued to decline (*2*). On page 617 of this issue, Juan-Jordá *et al.* (*5*) illustrate a similar contrast among ocean predators and introduce an approach for continuously monitoring their conservation statuses. The authors found that the situations for tunas and billfishes have improved over the past decade, but not those for sharks. This contrast owes partly to management, but biological and socioeconomic factors also cause fisheries to affect these species differently.

The International Union for Conservation of Nature (IUCN) (*6*) labels a species as critically endangered, endangered, or vulnerable on the basis of how much its population has declined over the past three generations or 10 years, whichever period is longer. If threats to a species are considered to be poorly understood or managed, then the IUCN applies these endangerment labels when there is a smaller population decline as a precaution. Juan-Jordá *et al.* built upon this classification system, known

[1]Department of Environmental Studies, University of Colorado Boulder, Boulder, CO, USA. [2]Center for Social and Environmental Futures, Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, CO, USA. [3]Department of Economics, University of Colorado Boulder, Boulder, CO, USA.
Email: matthew.g.burgess@colorado.edu

PHOTO: SANJIT DAS/PANOS FOR FT MAGAZINE



Power plant air pollution limits and job creation in coal-producing regions are key to India's clean-energy goals.

is to encourage investment in solar and wind power plants in the coal-producing regions. These energy infrastructure investments will also provide local job opportunities as the regions transition away from fossil fuel–based power generation in the long term (*8*).

India faces the twin challenges of mitigating carbon emissions and meeting an increasing energy demand. Several policies and regulations have been introduced to reduce overall energy demand and increase the supply for zero-carbon electricity. For instance, the government has undertaken large-scale procurement of energy-efficient home and office appliances to decrease their prices. It has also created a market for energy-saving certificates, where businesses saving more energy than their targets can sell the leftover credit to another company, creating a monetary incentive for businesses to meet energy efficiency targets (*9*). In addition, it has introduced competition through bidding by renewable energy companies, targets for each state to purchase renewable energy, exemptions for renewable energy from transmission charges (*10*), and transmission infrastructure for renewable energy (*11*). These policies have since led to some of the world's lowest solar and wind energy prices ($30 to $40 USD per megawatt-hour) (*12*). By 2021, India had become the world's fifth-largest solar power producer, with a capacity of 50 GW, and the fourth-largest wind power producer, with a capacity of 40 GW (*13*).

Although the carbon emissions and gross domestic product per capita of India are still less than half the global average, the country has ambitious plans for mitigating its carbon emissions. At the 26th United Nations Climate Change Conference held in 2021 in Glasgow, UK, the Indian government pledged a net-zero emissions target by 2070 and a near-term target of producing 50% of its electricity from renewable energy by 2030 (*14*). For India, the pursuit of climate and renewable energy targets is important, as is public health, employment, and energy affordability across regions and communities to ensure equitable growth. As India continues to develop its economy, balancing the near-term and long-term effects of its electricity sector policies, as well as their impact on social inequalities, will be critical to ensure a low-carbon transition that is green as well as just. ■

REFERENCES AND NOTES

1. R. R. Mohan *et al.*, "Greenhouse Gas Emission Estimates from the Energy Sector in India at the Sub-national Level (version 2.0)" (GHG Platform India, 2019).
2. M. Cropper *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2017936118 (2021).
3. S. Sengupta *et al.*, *Science* **378**, eabh1484 (2022).
4. J. E. Stiglitz, N. Stern, "Report of the High-Level Commission on Carbon Pricing" (Carbon Pricing Leadership Coalition, 2017); https://bit.ly/3TkiiLP.
5. R. Deshmukh, A. Phadke, D. S. Callaway, *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2008128118 (2021).
6. E. Ela *et al.*, *Electr. J.* **29**, 51 (2016).
7. "Analysis of Factors Impacting Retail Tariff and Measures to Address Them" (Forum of Regulators, 2021); https://bit.ly/3EYieNy.
8. S. Pai, H. Zerriffi, J. Jewell, J. Pathak, *Environ. Res. Lett.* **15**, 034065 (2020).
9. "Roadmap of Sustainable and Holistic Approach to National Energy Efficiency" (Bureau of Energy Efficiency, 2019); https://beeindia.gov.in/sites/default/files/Roshanee_print%20version%282%29.pdf.
10. "Tariff Policy" (Ministry of Power, 2016); https://bit.ly/3eJHX1q.
11. "Report on Green Energy Corridors – II: Part-A" (Power Grid Corporation of India Ltd., 2022); www.powergrid.in/sites/default/files/footer/smartgrid/Green%20Energy%20Corridor%202-Part%20A.pdf.
12. "Renewable Power Generation Costs in 2020" (International Renewable Energy Agency, 2021); https://bit.ly/3Teskhw.
13. "Renewable Capacity Statistics 2022" (International Renewable Energy Agency, 2022); www.irena.org/publications/2022/Apr/Renewable-Capacity-Statistics-2022.
14. Prime Minister's Office, "National Statement by Prime Minister Shri Narendra Modi at COP26 Summit in Glasgow" (Press Information Bureau of India, 2021); https://bit.ly/3goRWcS.

ACKNOWLEDGMENTS

The views of the authors are personal and do not represent the views of the Central Electricity Regulatory Commission.

10.1126/science.ade6040

MARINE CONSERVATION

Good and bad news for ocean predators

Some tunas and billfishes are recovering, but sharks continue to decline

By **Matthew G. Burgess**[1,2,3] *and* **Sarah L. Becker**[1,2]

As human population and economies have grown rapidly over the past 100 years, ecosystems worldwide have faced increasing pressure from overexploitation, habitat destruction, and other threats (*1*). In the oceans, roughly half of all commercially harvested fish and invertebrate stocks became overfished during the 20th century (*2*), and larger predators, such as billfishes and sharks, also dwindled (*3*). The 21st century has seen some marine fish and invertebrate stocks begin recovering owing to management efforts (*4*), whereas poorly managed stocks continued to decline (*2*). On page 617 of this issue, Juan-Jordá *et al.* (*5*) illustrate a similar contrast among ocean predators and introduce an approach for continuously monitoring their conservation statuses. The authors found that the situations for tunas and billfishes have improved over the past decade, but not those for sharks. This contrast owes partly to management, but biological and socioeconomic factors also cause fisheries to affect these species differently.

The International Union for Conservation of Nature (IUCN) (*6*) labels a species as critically endangered, endangered, or vulnerable on the basis of how much its population has declined over the past three generations or 10 years, whichever period is longer. If threats to a species are considered to be poorly understood or managed, then the IUCN applies these endangerment labels when there is a smaller population decline as a precaution. Juan-Jordá *et al.* built upon this classification system, known

[1]Department of Environmental Studies, University of Colorado Boulder, Boulder, CO, USA. [2]Center for Social and Environmental Futures, Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, CO, USA. [3]Department of Economics, University of Colorado Boulder, Boulder, CO, USA.
Email: matthew.g.burgess@colorado.edu

PHOTO: SANJIT DAS/PANOS FOR FT MAGAZINE

as "Criterion A," and developed indices for assessing the endangerment levels of seven tuna species, six billfish species, and five shark species. Their indices can be used to assess endangerment continuously in time, instead of being limited to fixed IUCN assessment intervals.

The indices define a species as being adequately managed if its mortality rate is less than the mortality rate that can sustain the maximum yield for fisheries. By this measure, the statuses of the tunas and billfishes have improved, on average, during the 2010s, and the mortality rates of several populations have returned to the levels that can support a maximum sustainable yield. By contrast, the statuses of sharks have continued to deteriorate on average during this period, and their mortality rates have remained well above the maximum sustainable rate. Juan-Jordá *et al.* attribute some of this contrast to the improved management of commercial fishing for tunas and billfishes, but not for sharks. For example, the International Commission for the Conservation of Atlantic Tunas has been setting and monitoring catch limits for tunas (*7*). However, the authors also highlight that other biological and fishery factors are needed to explain this difference between shark and tuna and billfish status, which is consistent with previous findings (*8*, *9*). For example, the differences in economic value and population growth rate and how each species is affected by fisheries directly and indirectly are important considerations.

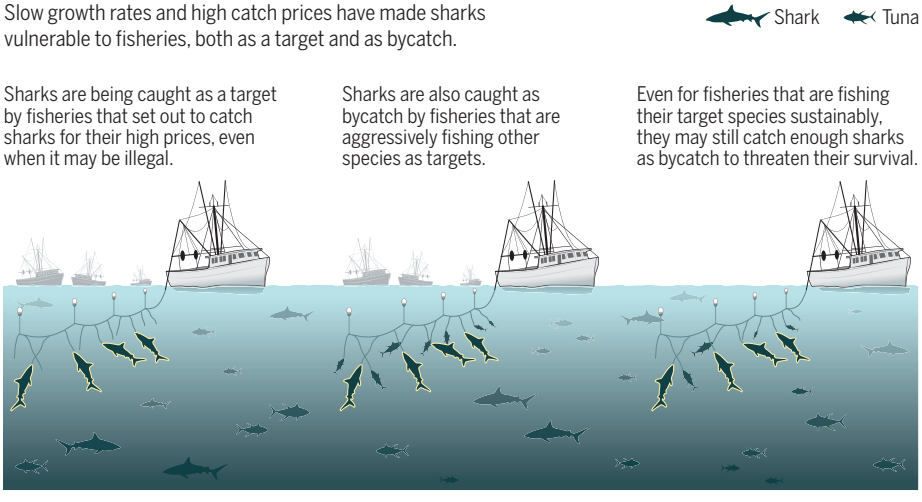
To understand why sharks are faring worse on average than tunas and billfishes, the mechanisms driving unsustainable fishing practices must be considered (see the figure). Without management (*4*), collective action, or community norms that promote cooperation (*10*), fisheries tend to overfish their target species. Commercially valuable species can support profitable fishing even at extremely low population sizes—if the species have high prices, large body sizes, low harvest costs, and/or small geographic ranges, which reduce the costs of catch (*11*). Nontarget species can also be affected by fishing activities, such as those that are caught unintentionally ("bycatch") (*12*) or opportunistically (for example, a fishing crew spotting and deciding to catch a different species than their original target) (*13*). Bycatch species can become threatened if they are frequently caught alongside overfished target species (*14*). They can also become threatened even if the target species are being sustainably caught when the bycatch species has a higher vulnerability—having a lower reproductive rate compared with its catch rate (*15*).

Some of the differences Juan-Jordá *et al.* found among sharks, tunas, and billfishes likely result from their different vulnerabilities to fishing activities. The five shark species studied by the authors all have slow population growth, have high vulnerability as bycatch, and are commonly caught by fisheries targeting tunas and billfishes. Sharks are also sometimes the target themselves. Although there has been some progress in managing fisheries that target sharks, these efforts face challenges posed by the lucrative fin trade and related illegal and unreported fishing (*9*). Marlins also stood out among the studied billfish species as being more endangered, likely because they are highly vulnerable as bycatch in tuna fisheries (*15*). By contrast, tuna species and relatively nonthreatened billfish species, such as swordfish, are mostly caught as targets (*7*). Among tuna species, their conservation statuses are more correlated with their biological and economic characteristics, such as short generation time and low price (which limit overfishing), than with the quality of their management (*8*).

Juan-Jordá *et al.* highlight the stark challenges facing oceanic predators—especially sharks. Successful shark conservation needs to address their specific biological and economic vulnerabilities, in addition to deploying fisheries management tools used for tunas and billfishes, such as science-based catch limits. Moreover, macroscopic ecosystem considerations may pose further challenges, even with well-managed predator fisheries. For example, maintaining sharks' ecosystem services as top predators might require higher shark abundances than is ideal for fishery catch. The conservation statuses of threatened target species can be improved by managing the fishing industry, which can benefit the industry economically in the long run while allowing the threatened species to recover (*2*, *14*). Generating sufficient scientific and governance capacity to implement successful management is often the primary challenge (*4*, *10*). However, the protection of high-vulnerability bycatch and nontarget species is expected to be more difficult because they will require fisheries to invest in better fishing gear and targeting practices, or reduce fishing efforts, without directly benefiting from these changes (*14*). The trade-offs between fishery benefits and ecosystem impacts will demand difficult negotiations and compromises between stakeholders. ■

How fisheries threaten sharks

Slow growth rates and high catch prices have made sharks vulnerable to fisheries, both as a target and as bycatch.



GRAPHIC: A. MASTIN/*SCIENCE*

REFERENCES AND NOTES

1. P. M. Vitousek, H. A. Mooney, J. Lubchenco, J. M. Melillo, *Science* **277**, 494 (1997).
2. C. Costello *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 5125 (2016).
3. D. J. McCauley *et al.*, *Science* **347**, 1255641 (2015).
4. R. Hilborn *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 2218 (2020).
5. M. J. Juan-Jordá *et al.*, *Science* **378**, eabj0211 (2022).
6. IUCN, *IUCN Red List Categories and Criteria: Version 3.1* (IUCN, ed. 2, 2012).
7. M. J. Juan-Jordá, H. Murua, H. Arrizabalaga, N. K. Dulvy, V. Restrepo, *Fish Fish.* **19**, 321 (2018).
8. M. Pons, M. C. Melnychuk, R. Hilborn, *Fish Fish.* **19**, 260 (2018).
9. N. K. Dulvy *et al.*, *Curr. Biol.* **27**, R565 (2017).
10. N. L. Gutiérrez, R. Hilborn, O. Defeo, *Nature* **470**, 386 (2011).
11. M. G. Burgess *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 3945 (2017).
12. R. Lewison, L. Crowder, A. Read, S. Freeman, *Trends Ecol. Evol.* **19**, 598 (2004).
13. T. A. Branch, A. S. Lobo, S. W. Purcell, *Trends Ecol. Evol.* **28**, 409 (2013).
14. M. G. Burgess *et al.*, *Science* **359**, 1255 (2018).
15. M. G. Burgess, S. Polasky, D. Tilman, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 15943 (2013).

ACKNOWLEDGMENTS

M.G.B. and S.L.B. thank C. Brooks, R. Langendorf, M. Hegwood, and N. O'Reilly for feedback.

10.1126/science.add0342

IMMUNOTHERAPY

Improving antitumor T cells

Disrupting cell cycle regulators can overcome anticancer T cell dysfunction

By **Caitlin C. Zebley**[1,2] *and* **Ben Youngblood**[2]

T cell–based immunotherapy strategies to treat cancer rely on the pool of tumor-specific T cells (engineered or endogenous) to retain a developmental potential that enables them to clonally expand and serially kill antigen-positive cells to achieve an efficacious immune response. Therefore, identifying T cell differentiation regulators in both clinical and preclinical settings is important to improve on current successes (*1*–*3*). Persistence of effector T cells has been identified as one of the critical barriers that limits responses to immunotherapies and is now the focus of current research efforts to sustain antitumor immunity (*4*, *5*). On page 616 of this issue, Freitas *et al.* (*6*) describe an unbiased genome-wide CRISPR-Cas9 screen to determine targets that enhance engineered chimeric antigen receptor (CAR) T cell effector function.

CAR T cells are a form of immunotherapy that are generated by ex vivo engineering a pool of T cells to express a specific receptor that recognizes cancer antigens; when infused into a patient, they elicit antitumor immune responses. Using a CRISPR screen with freshly isolated human T cells to identify genes that enhance CAR T cell effector function, Freitas *et al.* identified the genes mediator complex subunit 12 (*MED12*) and cyclin C (*CCNC*), which encode proteins in the cyclin-dependent kinase (CDK) module of the Mediator complex. The Mediator complex plays an integral role in transcription by acting as a bridge between transcription factors and the RNA polymerase machinery. The authors showed that genetic disruption of the CDK module of Mediator in human T cells resulted in enhanced effector function, metabolic fitness, and increased antitumor activity in several mouse tumor models treated with engineered CAR and T cell receptor (TCR) T cell immunotherapy.

To generate a quantity of CAR T cells large enough to provide therapeutic benefit for a patient, current production protocols often expand the cells ex vivo over the course of 1 to 2 weeks. However, this extended proliferation of T cells can come at the expense of developmental potential, which can limit the ability of the cells to undergo expansion after infusion into the patient. This quantity-versus-quality dilemma may be partially resolved by a better understanding of the relationship between cell cycle control and T cell terminal differentiation. Broadly, CDKs regulate cell division and RNA polymerase II–dependent transcription. CDK inhibitors have been shown to improve the effectiveness of immunotherapy for certain types of cancers (*7*, *8*). Freitas *et al.* found that disruption of the Mediator CDK module in human CAR T cells preserved the ability of T cells to mount a robust antitumor response after substantial ex vivo and in vivo expansion.

"...specific epigenetic programs reinforce the developmental transition of the T cell to a terminally differentiated state..."

In addition to improving the performance of engineered CAR and TCR T cells, these findings reveal an important aspect of T cell biology. Deletion of *MED12* resulted in enrichment of a T cell population with an effector-like phenotype that had reduced terminal differentiation. Mechanistically, disruption of MED12 altered the epigenetic landscape of T cells, resulting in increased chromatin accessibility at genomic regions enriched for binding motifs for the transcription factor families signal transducer and activator of transcription (STAT) and activating protein 1 (AP-1). Consequently, Mediator was able to bind to these epigenetically permissive regions and facilitate transcription of effector genes. The resulting epigenetically enhanced CAR T cells exhibit superior antitumor function. These data suggest that interruption of the CDK module through disruption of *MED12* or *CCNC* results in both epigenetic and transcriptional alterations that enhance CAR T cell effector potential.

Direct alteration of chromatin accessibility regulators has previously been used to improve CAR T cell function. Epigenetic-based strategies, such as the chemotherapeutic drug decitabine (which induces hypomethylation), inhibition of enhancer of zeste homolog 2 (EZH2, which blocks histone 3 Lys[27] trimethylation), and deletion of DNA methyltransferase 3α (*DNMT3A*), have been used to directly alter the epigenetic profile of ex vivo CAR T cells (*1*, *2*, *9*). From these studies, it has become clear that specific epigenetic programs reinforce the developmental transition of the T cell to a terminally differentiated state and that therapeutic efforts to block these epigenetic changes enhanced the durability of the CAR T cell–mediated antitumor response. Given that the CDK module of Mediator can control the G0-to-G1 transition, promoting entry into the cell division cycle, the findings by Freitas *et al.* suggest a relationship between cell cycle control and epigenetic programming and provide an indirect approach for reprogramming the epigenetic status of T cells used for immunotherapy.

Harnessing the robust killing potential of T cells through engineering efforts that refocus the effector response toward pathogenic cells (tumor cells and/or virally infected cells) has been revolutionary. Building on the success observed in cancer settings, the concept of CAR T cell therapy is now being applied to treat a range of diseases, such as cardiac fibrosis, as well as aging. CAR T cells targeting fibroblast activation protein (FAP) have successfully removed cardiac fibrosis in mouse models and improved cardiac function (*10*). Similarly, senolytic CAR T cells have been designed to target a protein expressed in most senescent cells (which lack the ability to proliferate). Infusing senolytic CAR T cells into mice with experimentally induced liver disease removed the senescent cells and reversed the disease phenotype (*11*). These studies highlight the potential for CAR T cell therapy to have a major effect on many aspects of human health and highlight the need for deeper understanding of the mechanisms currently limiting T cell proliferation and effector potential. ■

[1]Department of Bone Marrow Transplantation and Cellular Therapy, St. Jude Children's Research Hospital, Memphis, TN, USA. [2]Department of Immunology, St. Jude Children's Research Hospital, Memphis, TN, USA.
Email: benjamin.youngblood@stjude.org

REFERENCES AND NOTES

1. B. Prinzing *et al.*, *Sci. Transl. Med.* **13**, eabh0272 (2021).
2. E. W. Weber *et al.*, *Science* **372**, eaba1786 (2021).
3. M. Sadelain, I. Rivière, S. Riddell, *Nature* **545**, 423 (2017).
4. C. C. Zebley *et al.*, *Cell Rep.* **37**, 110079 (2021).
5. S. L. Maude *et al.*, *N. Engl. J. Med.* **371**, 1507 (2014).
6. K. A. Freitas *et al.*, *Science* **378**, eabn5647 (2022).
7. R. V. Uzhachenko *et al.*, *Cell Rep.* **35**, 108944 (2021).
8. P. K. Parua, R. P. Fisher, *Nat. Chem. Biol.* **16**, 716 (2020).
9. Y. Wang *et al.*, *Nat. Commun.* **12**, 409 (2021).
10. H. Aghajanian *et al.*, *Nature* **573**, 430 (2019).
11. C. Amor *et al.*, *Nature* **583**, 127 (2020).

ACKNOWLEDGMENTS

The authors are supported by the National Institutes of Health (R01CA237311 to B.Y.), a National Comprehensive Cancer Network Young Investigator Award (to C.C.Z.), an Alex's Lemonade Stand Young Investigator Grant (to C.C.Z.), Stand Up To Cancer (SU2C) (B.Y.), and the American Lebanese Syrian Associated Charities (B.Y. and C.C.Z.). C.C.Z. and B.Y. hold patents related to epigenetic biomarkers and methods for enhancing CAR T cell function.

10.1126/science.adf0546

HYPOTHESIS

Rewilding plant microbiomes

Microbiota of crop ancestors may offer a way to enhance sustainable food production

By **Jos M. Raaijmakers**[1,2] *and* **E. Toby Kiers**[3]

Over the past decade, research has shown that microorganisms living on and inside eukaryotes—the microbiota—are drivers of host health. For plants, microbiota can greatly expand their genomic capabilities by enhancing immunity, nutrient acquisition, and tolerance to environmental stresses (*1*). More than ever, plant microbiota are being considered as a lever to increase the sustainability of food production under a changing climate. Emerging from this global interest to harness the largely unexplored functional potential of microbiota, the microbiome rewilding hypothesis posits that plant and animal health can be improved by reinstating key members of the diverse (ancestral) microbiota that were lost through domestication and industrialization processes, including changes in diet, plant and animal breeding, and the (over)use of antibiotics, pesticides, and fertilizers (*2*–*4*).

A central question is whether the microbiomes of crop ancestors can be used to "rewild" microbiomes of current crops. Similar to reversing industrialization-associated changes in human gut microbiota (*5*), plant microbiome rewilding builds on the premise that wild ancestors harbor microbial genera with specific traits that are not found (or are strongly depleted) in the microbiome of modern crops. To date, however, it is unknown for most plant species whether (and which) microbial genera and functions were lost during plant domestication, and to what extent rewilding can enhance the health and sustainability of modern crops. In animal systems, the effectiveness of rewilding approaches is intensely debated (*3*), and similar discussions are needed for crop rewilding approaches.

Plant domestication is one of the most important accomplishments in human history, helping drive the transition from a nomadic to a sedentary lifestyle. Through stepwise processes, crop plants acquired a suite of new traits, including larger seeds, determinate growth, photoperiod sensitivity, and reduced levels of bitter substances. Although this led to a more continuous food supply, domestication caused a reduction in plant genetic diversity because only desired alleles were spread, while genomic regions next to the target genes suffered selective sweeps (*6*). This so-called "domestication syndrome" decreased the ability of crops to withstand pests and diseases.

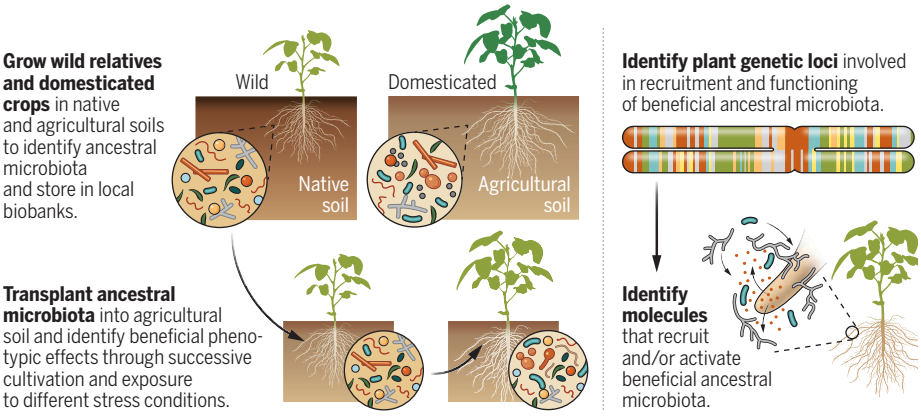
Domestication was also accompanied by considerable habitat expansion and management practices, with increased reliance on external inputs (pesticides, fertilizers, fresh water) to obtain higher yields and to protect domesticated crops from abiotic and biotic stresses. This transition of plants from their native habitats to physicochemically diverse agricultural soils also led to substantial changes in microbiome composition (*7*). For example, the domestication of legumes, combined with long-term nitrogen fertilization, has been linked to the evolution of less mutualistic rhizobia (nitrogen-fixing soil bacteria that form nodules in the roots of legumes), and legume varieties that are less able to discriminate between rhizobia that provide nitrogen to the plant versus those that do not (*4*, *8*). Disruption of the symbiotic interaction between plants and mutualistic root fungi, known as mycorrhizae, has also been documented, with domesticated crops showing lower colonization and a decreased growth response to fungal symbionts, especially in fertilized soils (*9*). Alterations in the genetic makeup of crops could at least partially explain this altered assembly of microbial symbionts.

In theory, rewilding can be achieved by following a simple roadmap that identifies beneficial microbiota from wild progenitors in their sites of origin, experimentally quantifies their beneficial effects on modern crops over multiple generations, and disentangles the plant genetic basis of microbial colonization (see the figure). However, it is unknown whether domestication and industrialization have affected microbiome assembly in consistent and convergent ways across diverse crop lineages. This is important because if there is predictability in assembly processes, rewilding approaches are likely to be more successful.

Microbiome assembly encompasses three types of interactions—host-to-microbe, microbe-to-host, and microbe-to-microbe—each with its own evolutionary features (*10*). Most well-studied are the selective pressures by which hosts select for beneficial interactions with their microbiome. This means hosts can modulate physical and chemical conditions to control an ever-evolving microbial community (*10*). For example, plants actively release exudates, such as sugars, amino acids, and volatile and secondary metabolites, that attract and selectively enrich for specific microbial species (*1*). These regulators impose selection pressures on microbiome assembly that can persist over multiple plant generations, leading to differential effects on host performance (*11*). Although host-mediated recruitment and selection of

A research roadmap to microbiome-assisted crops

Rewilding involves reinstating key ancestral microbes in agricultural soils or planting materials, and/or breeding modern crops with specific traits that support ancestral microbiota colonization. By comparing wild and domesticated plants, beneficial ancestral microbiomes can be characterized to make rewilding a possibility.



GRAPHIC: K. HOLOSKI/SCIENCE

[1]Netherlands Institute of Ecology, Wageningen, Netherlands. [2]Institute of Biology, Leiden University, Leiden, Netherlands. [3]Amsterdam Institute for Life and Environment, Vrije Universiteit Amsterdam, Amsterdam, Netherlands. Email: j.raaijmakers@nioo.knaw.nl

microbes have only been shown for relatively few exudate compounds (e.g., strigolactones, flavonoids, coumarins, benzoxazinoids, and organic acids), there is likely a large repertoire of unknown signals used by wild crop ancestors to control the composition and activity of their microbiomes.

The host and microbiota, however, are not a single evolutionary unit acting with a common interest (*10*). Instead, microbes evolve strategies to maximize their own success, driven by strong selection pressures to compete and persist on or within a host plant, even if there are costs to the host (*12*). In response, plants have evolved mechanisms to monitor and control the reproductive success of microbial partners. Because plant domestication is a process of artificial selection and not intentionally directed toward maximizing positive microbe-to-host effects, these host control mechanisms may have diminished. Indeed, studies with wheat, maize, and soybean show a reduction in positive microbe-to-host effects following domestication (*4*). If selection to control microbes decreases, rewilding strategies may become even more challenging because host plants will have difficulty establishing mutualistic interactions.

Identifying differences in the microbiomes between crop plants and their wild progenitors—and the mechanisms mediating these differences—is straightforward in theory, but difficult in practice. Tools such as high-throughput plant phenotyping and genotyping, next-generation sequencing, and advanced metabolomics can provide insights into the genetic and chemical basis of the domestication syndrome. To date, however, relatively few plant traits have been attributed to single genes or pathways. On the microbiome side, it is likewise challenging to identify specific genes involved in host selection. Large-scale comparative (meta)genomics analyses are starting to reveal genes involved in microbial adaptation to the plant environment (*13*). These approaches can help drive microbiome rewilding by identifying key genes underlying beneficial interactions across crop species.

To unravel plant-specific mechanisms involved in microbiome assembly, creating segregating populations by crossing the crop cultivar and its wild progenitor is one of the most comprehensive approaches to date. Recombinant inbred lines have already been instrumental in identifying genetic loci that control various agronomically relevant traits, such as pathogen resistance and drought tolerance, and are now also being adopted to map the microbiome as a plant phenotype (*14*). Through genome-wide association mapping, quantitative trait analyses, and fine mapping, domestication loci associated with microbiome assembly could be identified. However, validation of candidate plant and microbiome genes is still challenging because it requires incorporating environmental effects on gene expression.

To understand, and ultimately implement, microbiome rewilding in future crop production, biogeographical analyses of native soils in a crop's center of origin are needed. This can be combined with profiling the microbiomes of wild relatives, landraces (domesticated, locally adapted variety of a plant species), and improved crop varieties growing in native and agricultural soils. These data will help to interpret shifts in plant microbiome composition along the domestication path, and to disentangle the importance of plant-soil feedback in driving these shifts. Considering that the effects of plant genotype on microbiome composition may be relatively small, exposure of crops to a suite of different stresses can help amplify the differential recruitment of specific ancestral microbiota. Preservation and genomic characterization of microbiomes (i.e., biobanking) of wild crop progenitors from their centers of origin should also be prioritized.

Additionally, microbiome transplants should be integrated in experiments to pinpoint the key beneficial ancestral microbiota. Analogous to fecal microbiota transplantation to redirect the dysbiotic composition of human microbiomes, transferring complex microbial communities from root or shoot tissue of wild crop relatives onto seeds or planting material of their domesticated counterparts can initially be used to identify specific microbiome-associated plant phenotypes. If beneficial effects of wild microbiome transplants on crop cultivars are confirmed, approaches to minimize microbiome complexity through dilution to extinction or the design of synthetic microbial communities may help identify the key microbial genera associated with particular plant phenotypes.

The performance of modern crop varieties may also be improved through artificial selection on the native microbiome. In this experimental evolution approach, the microbiome as a whole is propagated by experimental passage to the next generation of hosts, while the host is kept genetically invariant (*15*). Microbiome functions that have been artificially selected can then be identified by comparative taxonomic and functional profiling of the communities that evolved under different selection regimes. Exposing plants to specific stresses, such as drought or nutrient deprivation, can expedite the search for specific subsets of ancestral beneficial consortia. This can be followed with the identification of specific plant root exudates (or other control mechanisms) in wild relatives that are responsible for the recruitment of these consortia. Existing domesticated crop varieties can then be selected for those particular control mechanisms and used as hosts compatible with the ancestral consortia.

Another approach involves identifying corresponding loci from wild crop ancestors and reintroducing these in domesticated crops, which is conceptually similar to the transfer of disease resistance genes lost during domestication. However, the success of this approach is questionable considering that the host traits involved in microbiome assembly can be complex and multifactorial, which may not work in a predictable way. Instead of transferring complex traits from wild ancestors to domesticated crops, one may consider de novo domestication—the concept that domestication genes are introduced into the wild crop relatives, rather than vice versa (*6*).

Rewilding approaches can offer a new avenue to harness the benefits of ancestral microbiota, and do not preclude the use of domesticated crop cultivars or agricultural management practices, such as fertilizer. Integrating field biology experiments with reductionist approaches in controlled conditions will be instrumental in defining the genetic and chemical basis of the diverse services that microbes can offer crops. Of the steps needed to accomplish tangible results, rigorous profiling and biobanking of microbiomes in their centers of origin and designing compatible combinations of host plant and beneficial ancestral microbiota are probably the most challenging. As rewilding research moves between the field and the lab, its value and integration in breeding programs for a new generation of "microbiome-assisted" crops await critical assessment in different agroecologies. ■

REFERENCES AND NOTES

1. P. Trivedi, J. E. Leach, S. G. Tringe, T. Sa, B. K. Singh, *Nat. Rev. Microbiol.* **18**, 607 (2020).
2. M. G. Dominguez-Bello *et al.*, *Science* **362**, 6410 (2018).
3. A. T. Reese *et al.*, *eLife* **10**, e60197 (2021).
4. S. S. Porter, J. L. Sachs, *Trends Ecol. Evol.* **35**, 426 (2020).
5. R. N. Carmody, A. Sarkar, A. T. Reese, *Science* **372**, 462 (2021).
6. A. R. Fernie, J. Yan, *Mol. Plant* **12**, 615 (2019).
7. J. E. Pérez-Jaramillo, V. J. Carrión, M. de Hollander, J. M. Raaijmakers, *Microbiome* **6**, 143 (2018).
8. J. Liu, X. Yu, Q. Qin, R. D. Dinkins, H. Zhu, *Front. Genet.* **11**, 00973 (2020).
9. N. Martín-Robles *et al.*, *New Phytol.* **218**, 322 (2018).
10. K. R. Foster, J. Schluter, K. Z. Coyte, S. Rakoff-Nahoum, *Nature* **548**, 43 (2017).
11. P. A. H. M. Bakker, C. M. J. Pieterse, R. de Jonge, R. L. Berendsen, *Cell* **172**, 1178 (2018).
12. M. Klein *et al.*, *Evol. Appl.* (2021).
13. A. Levy *et al.*, *Nat. Genet.* **50**, 138 (2017).
14. J. Bergelson, B. Brachi, F. Roux, F. Vailleau, *Curr. Opin. Biotechnol.* **70**, 167 (2021).
15. U. G. Mueller, J. L. Sachs, *Trends Microbiol.* **23**, 606 (2015).

ACKNOWLEDGMENTS

Thanks to D. Ramirez-Villacis for help with the figure. Support was provided by the Dutch Research Council (NWO)–Gravitation program Microp 024.004.014 (J.M.R., E.T.K.), NWO-Vici (E.T.K.), and Human Frontier Science Program RGP 0029 (E.T.K.).

10.1126/science.abn6350



POLICY FORUM

Plans for the UK's Drax power station include large-scale bioenergy with carbon capture and storage.

ENERGY AND CLIMATE

Industrial clusters for deep decarbonization

Net-zero megaprojects in the UK offer promise and lessons

By **Benjamin K. Sovacool[1,2,3], Frank W. Geels[4], Marfuga Iskandarova[3]**

Perhaps no sector of the global economy is in greater need of concerted efforts toward deep decarbonization than industry, which includes energy-intensive sectors such as chemicals, iron and steel, cement, and aluminum (*1*). Yet industry has long been perceived as hard to decarbonize and has been mostly sheltered from strong energy and climate policies over concerns about potential job losses, national competitiveness, and carbon leakage. Industrial decarbonization scenarios often identify carbon capture and storage (CCS) and fuel switching to hydrogen as potential net-zero options (*2*), but these technologies are expensive for individual companies and specific industries. These options can become more feasible when implemented in industrial clusters, where plants from different industries operate in close proximity. We see promise and lessons in recent advancements in the coevolution of net-zero cluster planning, policy implementation, and technical development in the UK, where world-leading plans and designs have progressed close to the implementation stage.

Despite considerable growth in production and energy use, the industrial sector's energy mix has remained virtually unchanged, remaining heavily connected to fossil fuels, especially coal. Whatever metric is used, industrial emissions have grown faster than any other sector, driven by both increased materials and mineral extraction, as well as higher rates of manufacturing and production (*3*). Recently, however, net-zero commitments have been leading to increased policy interest in industrial decarbonization. Policy-makers must find leverage points for reaching net-zero emissions while also allowing industries to continue to grow and prices for products to remain affordable (*4*).

Despite the potential of technologies such as CCS, many obstacles remain. For example, their scale is small; industry as a whole captures only 40 million metric tons of carbon dioxide (CO_2) emissions per year, and adding all planned carbon capture units to the tally increases that number only to 140 million metric tons (*5*). The motivation of a cluster approach is that the construction of cluster-wide CCS infrastructures can enable the transport of CO_2 from multiple plants and storage in offshore saline aquifers or empty gas fields as well as the creation and use of low-carbon sources such as hydrogen, mainly "blue" hydrogen produced from natural gas.

NET-ZERO INDUSTRY POLICY IN THE UK

UK policy-makers started exploring the industrial decarbonization challenge with a series of roadmaps, action plans, and strategies developed with industry from 2015 to 2017. These highlighted the need to go beyond energy and material efficiency innovations toward fuel switching and CCS. The government's 2018 CCS Action Plan emphasized an industrial cluster approach and articulated targets such as the development of the first carbon capture, utilization, and sequestration (CCUS) facility by the mid-2020s and deployment at scale during the 2030s. In 2018, policy-makers created the Industrial Strategy Challenge Fund, which allocated £170 million to clean growth and transforming construction and was in turn matched by £250 million of private sector investment. This fund enabled industrial firms to engage in more detailed pre–front-end engineering and de-

[1]Department of Earth and Environment, Boston University, Boston, MA, USA. [2]Center for Energy Technologies, Department of Business Development and Technology, Aarhus University, Aarhus, Denmark. [3]Science Policy Research Unit (SPRU), University of Sussex, Falmer, UK. [4]Manchester Institute of Innovation Research, University of Manchester, Manchester, UK. Email: b.sovacool@sussex.ac.uk

PHOTO: DAN KITWOOD/GETTY IMAGES

sign (FEED) studies of low-carbon technologies (*6*).

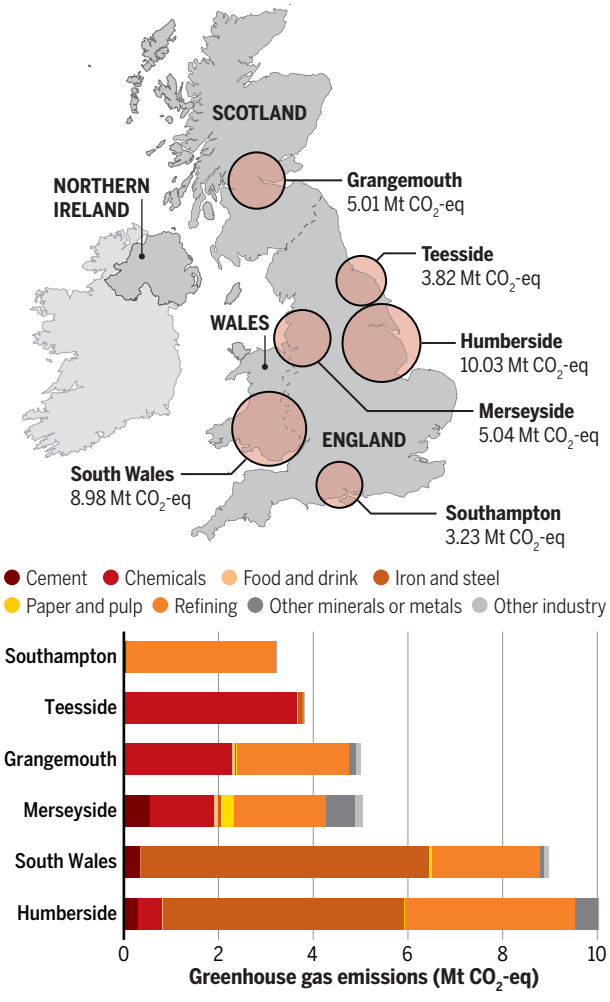
Policy momentum increased in 2019, when the UK government enshrined a net-zero emission target in law by amending the 2008 Climate Change Act, and in 2020, when the Ten Point Plan for a Green Industrial Revolution aimed for the production and use of 5 GW of low-carbon hydrogen by 2030 (mostly from natural gas and CCS) and the deployment of CCS in two industrial clusters by 2025 and four clusters by 2030. UK ambition has recently been doubled to up to 10 GW of low-carbon hydrogen production capacity by 2030, with at least half of this being from electrolytic hydrogen that is intended to be sourced from future renewable energy and nuclear power capacity (*7*).

To support the deployment of these technologies in the coming years, policy-makers created a £240 million Net Zero Hydrogen Fund and a £1 billion CCS Infrastructure Fund, which are being implemented through a cluster-sequencing strategy in which cluster-based partnerships first apply for funding for transport and storage infrastructures and then, in a second phase, for funding for CO_2 capture installations, hydrogen production, and fuel switching. To bypass previously documented difficulties with piloting and deploying carbon capture technologies, such as those that occurred with the $7.5 billion Kemper project in the United States, the UK is entirely avoiding the use of coal, is avoiding troublesome technical features such as integrated gasification combined-cycle turbines, and is not seeking precombustion capture at large power plants, related in part to earlier failures in government support and policy for CCS.

The UK's net-zero cluster plans have several challenging characteristics of "megaprojects" (*8*) such as high technological novelty (because many of the technological components are new to the UK or have not yet been applied at the scale involved here), high structural complexity (because multiple components need to be integrated into a working system), and high pace (to meet government targets for 2025 and 2030). These net-zero megaprojects offer a rare opportunity to identify salient real-world implementation challenges with exceptional potential to inform global efforts at decarbonization elsewhere.

Decarbonization in UK industrial clusters

The map shows the location of the six largest clusters in terms of annual greenhouse gas (GHG) emissions (million metric tons (Mt) of CO_2-eq) (top). The graph shows emissions from different industrial sectors (excluding power generation) in six UK clusters (bottom), dominated mostly by chemicals, refining, and iron and steel. Each of these clusters has aggressive plans in place for deployment of net-zero infrastructure.



FROM INDIVIDUAL PROJECTS TO INTEGRATED CLUSTERS

UK industrial decarbonization policy has moved relatively quickly from problem exploration to general vision to implementation and is increasingly focused on industrial clusters: large-scale facilities for colocated energy production, industrial manufacturing, distribution, and transportation (*9*). Six specific clusters, which account for more than 50% of direct carbon emissions from industry (*10*), are spread across Scotland (Grangemouth), Wales (South Wales), and England (the Humber, Merseyside, Southampton, and Teesside) (see the figure).

Although these clusters have different emission profiles, the three most important industries are oil refining, chemicals, and iron and steel (see the figure). Five clusters submitted proposals to the first round of the CCS Infrastructure Fund. In October 2021, the government selected the HyNet project (in Merseyside) and the East Coast Cluster (which includes both Teesside and the Humber clusters) as potential first locations for CO_2 transport and storage networks in the mid-2020s. In August 2022, the government shortlisted 20 power CCUS, hydrogen, and industrial carbon capture projects in these two clusters to proceed to the due diligence stage of the phase 2 cluster-sequencing process.

These UK efforts at industrial decarbonization did not arise in a vacuum but rather emerged in a regional context where Dutch, Danish, and Norwegian efforts have shaped the UK approach. In Denmark, the Kalundborg Eco-Industrial Park has been operating since the late 1960s and has sought to achieve industrial symbiosis by integrating electricity supply, heat production and distribution, steam generation for oil refining, and the use of surplus heat in homes and commercial enterprises. In Norway, Shell, Total, and Equinor launched the Northern Lights project in 2020, with completion due to occur in 2024, seeking to establish the first cross-border carbon transport and storage network in the world. The Port of Rotterdam in the Netherlands launched its Porthos (Port of Rotterdam CO_2 Transport Hub and Offshore Storage) initiative in 2022 to store ~2.5 million metric tons of CO_2 per year offshore in the North Sea, ~20 km off the coast.

None of these efforts, however, will operate at the larger scale of activities in the UK, with the Kalundborg project representing an integration of industrial activities but not decarbonization, the Northern Lights project representing only a transport and storage network, and the Porthos initiative erecting only a storage site. UK plans seek to integrate multiple technologies over industrial clusters that are much larger in terms of volume of manufacturing and production and greater in terms of the carbon emissions to be captured than those of its European neighbors.

Technological challenges

The net-zero cluster plans are technologically complex megaprojects because they involve many innovative technologies that need to be integrated into wider systems. They are also encountering some emergent innovation barriers such as solvents for CCS, technical standards and specifications for hydro-

CREDIT (GRAPHIC) K. FRANKLIN/*SCIENCE*; (DATA) SEE SUPPLEM

gen pipelines, and fuel-blending challenges with hydrogen boilers. Net-zero plans in the Humber, for instance, involve new infrastructures such as large-scale bioenergy with CCS (BECCS) at Drax, new onshore pipelines for hydrogen and CO_2 transport to and from some of the main industrial emitters, offshore CO_2 pipelines, and CO_2 storage in the offshore Endurance saline aquifer. These plans additionally include CO_2 capture projects, blue hydrogen production projects (at Saltend Chemicals Park), green hydrogen production with electrolyzers (at Immingham), fuel-switching projects from natural gas to hydrogen (at the power plant of the chemicals park and the two Immingham refineries), and hydrogen storage projects (in onshore salt caverns or offshore empty gas fields) to manage demand fluctuations.

The Humber decarbonization plans are not only expensive, with combined capital costs running in the tens of billions of pounds, but also challenging because they involve the integration of many new technologies (e.g., BECCS, hydrogen, CCUS, and transport and storage networks). Firms and policy-makers therefore need to accommodate flexibility and learning by doing while simultaneously driving rapid implementation.

Firms intend to address this system integration and technological novelty challenge by using H2H Saltend as a "kickstarter project" that initially involves the production of blue hydrogen by Equinor (*11*), the use of 30% blended hydrogen in an adjusted gas turbine by Triton Power at the same park, and the construction of a CO_2 pipeline to Easington and then offshore. This relatively low-risk project enables learning about essential technical components while also constructing basic infrastructures. Learning experiences will then inform subsequent pipeline, CO_2 capture, and fuel-switching projects as they spread beyond Saltend.

Firms within the Humber and HyNet (Merseyside) also try to address the technological novelty challenge with pilot programs. Working with General Electric and Mitsubishi, they will test hydrogen cofiring in their combined-cycle gas turbines. The intent is to begin with a 30% hydrogen blend in 2023 and then move up to a 60% blend in 2025 before converting turbines to run entirely on hydrogen. UK firms are also working with Royal Dutch Shell and Equinor on techniques for pre- and postcombustion capture of the resulting carbon emissions, underscoring an international dimension of learning and technology transfer.

Policy-makers aim to address these challenges through their cluster-sequencing strategy, which first provides funding for infrastructure creation and then for a few initial CO_2 capture and fuel-switching projects. This modular approach not only alleviates chicken-and-egg coordination problems but also allows later CO_2 capture and fuel-switching projects to learn from earlier ones.

Organizational and policy challenges

Net-zero cluster plans are organizationally complex because no single organization is in charge. Instead, different (coalitions of) organizations intend to implement different subprojects. In the Humber, for instance, National Grid will build and operate the onshore hydrogen and CO_2 pipelines; the Northern Endurance Partnership, which includes BP, Equinor, Shell, Total, and National Grid, will build and operate the offshore CO_2 pipelines and storage; and industrial emitters will oversee their own CO_2 capture or fuel-switching projects, sometimes together with co-located firms or suppliers.

This decentralized organizational structure means that the megaproject can be implemented in a modular fashion, which increases flexibility but can also generate coordination and system integration problems. Pipeline constructors, for instance, will not make a final investment decision until they have contractual certainty that industrial emitters will transport certain amounts of CO_2, for which they will pay the pipeline operators. But industrial emitters will not make such decisions until they have more certainty about policy support and institutions, which have not yet stabilized. To improve cross-project coordination and prevent potential delays due to "waiting games," there are cluster-wide meetings and platforms to exchange information and facilitate discussions, but these cannot avoid disputes, which are inevitable as a result of sometimes conflicting interests and political agendas (*12*, *13*).

Net-zero cluster plans are also institutionally complex because policies and institutions are essential drivers that are still under development. Policy goals and strategies evolved rapidly in the past few years, which increased stakeholder confidence about the direction of travel. But firms also need more clarity about implementation-oriented financial support schemes and operational business models to move from FEED studies to final investment decisions. Without substantial public support for capital expenditures (especially in infrastructures) and operational expenditures, most industrial firms facing cutthroat international competition are unlikely to invest in low-carbon technologies.

An implementation challenge is to design policies that accommodate substantial techno-economic differences between industrial firms while preventing the creation of an overly complex policy landscape. To navigate this challenge, UK policy-makers interacted with firms through responses to government consultations, expert groups, and bilateral discussions over the past few years, leading to substantial policy learning and adjustments, which have started to converge on several business models and policy interventions (see the box). Many of these business models and policy mechanisms are innovative themselves, and some, such as capacity payments to power producers, were previously overruled in several countries as a way to protect old thermal power plants.

Although incentives aim to provide longer-term financial certainty, they are not yet finalized, which means that firms cannot complete the business case for their investment decisions. Another complication is that the development of regulations and standards (e.g., CO_2 footprint and purity of low-carbon hydrogen) has received less attention and is lagging behind. Because these standards affect technical designs and costs, their underdeveloped status may also cause delays. The alignment between different business models has also not yet been fully considered, which may cause problems for later system integration of different technical modules.

A final area where UK policy-makers are seeking to overcome prospective challenges relates to broader social legitimacy and

Business models, policy mechanisms

1. **Dispatchable power agreements**
   Offer power plants with carbon capture and storage a payment for available capacity and a variable payment per megawatt-hour of generated low-carbon electricity.
2. **Industrial carbon capture business models**
   Incentivize the deployment of carbon capture technology by industrial users with industrial carbon capture contracts to provide ongoing revenue support and capital grant funding where relevant.
3. **Low-carbon hydrogen agreements**
   Pay hydrogen producers a flat (indexed) rate between the "strike price" (a price for electricity that reflects the investment cost for low-carbon technology) and the "reference price" (a measure of the average market price for electricity in the market).
4. **Regulated asset base model**
   Used for transport and storage infrastructures, regulated asset base models include a payment for the amount of CO_2 that has been moved and stored and a payment for building the infrastructure.
5. **Carbon border-tariff adjustments**
   Restrictions are placed on traded and imported carbon-intensive goods, which reduces leakage and ensures that carbon is more properly valued in the market.

public acceptance. Local concerns include employment opportunities, rights of way for new infrastructure, and the long-term environmental sustainability of using natural gas for blue hydrogen owing to fugitive emissions. Frequent public consultations have occurred across all six of the clusters and have been especially visible in places like the Humber. There, the city of Hull has run a campaign with regional partners (called "Oh Yes! Net Zero") to increase public knowledge about net-zero projects and inform and educate the public about skills and educational opportunities.

TENSIONS AND LESSONS

This industrial net-zero pathway involves the co-construction of new technologies, social networks, and institutions, and it enables learning processes that aim to address technological novelty and system-integration challenges. This sociotechnical transition process is coevolutionary and full of uncertainties, negotiations, and implementation struggles (*14*).

Although the UK's net-zero cluster plans have progressed well, they also exhibit distinct tensions. One tension exists between the desire for high-speed delivery to meet targets and allowing for flexibility and learning, which is needed because there are many new components that have never before been integrated at the ambitious scale that is required in the UK. Another tension exists around the modular approach that the UK cluster-sequencing strategy has used by separately advancing and funding infrastructures and plant-specific carbon capture and hydrogen projects. This has enabled progress but may not yet pay sufficient attention to system integration issues. Megaprojects with many innovative components usually go through several rounds of learning and design adjustments, which has not happened in the UK, where the hope is that everything will fall into place.

Other risks exist not only in isolation but also in relation to each other. One of these is skills development, which includes the possible lack of skills or construction labor that the UK may have if it embarks on decarbonization across multiple clusters simultaneously. The HyNet Consortium based at the University of Chester's Thornton Science Park is being funded by the government to "pioneer the new skills required to meet the country's Net Zero targets," but it will support only one cluster, at Merseyside. Another risk is associated with increasing interdependence between industries. The closure or withdrawal of a major firm, such as the Essar Refinery in Merseyside or Equinor or Drax in the Humber, could negatively affect other companies and the viability of net-zero projects.

Despite these tensions and risks, the UK experience points the way toward an approach that enables collective action between companies in a cluster to build new systems together, with a particular emphasis on overcoming technological, organizational, policy, and even social implementation challenges. UK policy-makers have pioneered new policy mechanisms that could offer a template for global industrial decarbonization, such as a cluster-sequencing strategy that enables a modular approach to megaproject development and differentiated business models to accommodate industry specificities and different system components. Interactions with firms also enable policy-makers to learn and adjust policy designs over time while also increasing stakeholder confidence in the government's willingness to drive and support industrial decarbonization.

Salient domestic and international policy implications arise as well. In the UK, most policy attention has focused on technologies and finance, which are essential prerequisites for many industrial organizations. Less attention has been given to wider sociotechnical system issues such as skills (including welders, machinists, and civil engineers) or standards and regulations (e.g., CO_2 specifications, health and safety regulations, planning requirements), which could lead to delays. Skills development can be addressed by further investments in infrastructure, such as the HyNet Academy, across all clusters, and standards and regulations can be tackled by proactive efforts within government. Internationally, the UK is trialing new forms of policy intervention such as regulated asset base models, differentiated strike prices for carbon, and cross-border tariffs, which all seek to stimulate a market for local decarbonization and attempt to minimize carbon leakage.

Nevertheless, not every aspect of the UK approach will be generalizable to other countries. Some elements will be difficult to emulate, such as the smaller nature of the UK economy compared with the more spatially and industrially diverse economies of China and the United States; the binding and stringently enforceable nature of the Climate Change Act, which is (so far) distinctive to the UK; the availability of adequate carbon storage capacity in both depleted oil and gas reservoirs and salt caverns; and a readily available supply of natural gas to produce blue hydrogen. Policymakers around the world, however, can replicate other aspects of the UK approach, such as strong political commitment for industrial decarbonization across the political divide; consistent core climate policy and binding carbon targets that do not change unexpectedly over time; the importance of financing and stimulating decarbonization with public procurement and public revenue; and reliance on harbors, seaports, and pipelines to distribute hydrogen and transport CO_2 to remote areas and vital industrial clusters.

Ultimately, there are still uncertainties about financial support, standards, and system integration of innovative technologies. Much could still go wrong, the UK strategy remains centered on gas (and a near-term commitment to blue hydrogen, rather than green hydrogen, which implies environmental risks), and both hydrogen and CCS have suffered from media-hype cycles and a failure to deliver emissions reductions before (*15*). But the rapid development of technical designs, projects, coalitions, and policies has generated substantial momentum in the UK, with companies hoping to make final investment decisions in 2023 with the goal of delivering the first low-carbon industrial clusters by 2025 to 2027. Hopefully this momentum can encourage other industrial firms to follow the UK's lead, showing how such firms that bear considerable historical responsibility for climate change can harness the capabilities necessary for solving it. ■

REFERENCES AND NOTES

1. International Energy Agency (IEA), "Industry" (IEA, 2022); https://www.iea.org/reports/industry.
2. Committee on Climate Change, "Net zero technical report" (Committee on Climate Change, 2019).
3. Intergovernmental Panel on Climate Change, "Climate change 2022: Mitigation of climate change. Contribution of Working Group III to the sixth assessment report of the Intergovernmental Panel on Climate Change," P. R. Shukla *et al*., Eds. (Cambridge Univ. Press, 2022).
4. P. Fennell, J. Driver, C. Bataille, S. J. Davis, *Nature* **603**, 574 (2022).
5. R. F. Service, *Science* **371**, 1300 (2021).
6. UK Research and Innovation, "Industrial Strategy Challenge Fund" (2022); https://www.ukri.org/what-we-offer/our-main-funds/industrial-strategy-challenge-fund/.
7. Department for Business, Energy, and Industrial Strategy, "British energy security strategy (updated 7 April 2022)" (UK government policy paper, 2022).
8. B. Flyvbjerg, *Proj. Manage. J.* **45**, 6 (2014).
9. P. Devine-Wright, *Energy Res. Soc. Sci*. **91**, 102725 (2022).
10. HM Government, "Industrial decarbonisation strategy" (UK government policy paper CP 399, 2021).
11. Equinor, "H2H Saltend: The first step to a zero carbon Humber" (2020); https://www.equinor.com/energy/h2h-saltend.
12. M. Ringel, N. Bruch, M. Knodt, *Energy Res. Soc. Sci*. **77**, 102083 (2021).
13. J. Meckling, *Glob. Environ. Polit*. **21**, 134 (2021).
14. B. K. Sovacool, F. W. Geels, *Environ. Innov. Soc. Transit*. **41**, 89 (2021).
15. A. Martínez Arranz, *Glob. Environ. Change* **41**, 124 (2016).

ACKNOWLEDGMENTS

We acknowledge support by the UK Research and Innovation (UKRI) Industrial Strategy Challenge Fund (ISCF) Industrial Challenge as part of UK Industrial Decarbonisation Research and Innovation Centre (IDRIC) award number EP/V027050/1.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.add0402

10.1126/science.add0402



BOOKS *et al.*

Perception, movement, emotion, and cognition are irreducibly intertwined, argues a neuroscientist.

NEUROSCIENCE

Transcending reductionism in neuroscience

The brain is a relational organ that is not just the sum of its parts

By **Alex Gomez-Marin**

In his new book, *The Entangled Brain: How Perception, Cognition, and Emotion Are Woven Together*, Brazilian neuroscientist Luiz Pessoa offers a way to construe the brain as a fully integrated organ, a framework that "while not rare, is also *not* mainstream among neuroscientists." A "divide-and-conquer strategy" has produced ever more refined brain maps, he argues, and subsequent leaps from structure to function. However, not only are anatomical brain areas far from simply located units of cognition but, as the subtitle of the book makes explicit, perception, cognition, and emotion are also interweaved.

To stress the networked nature of the brain, Pessoa has chosen a timely adjective: "entangled." He seeks a portrait of the brain beyond the high-resolution caricature of cognitive functions placed inside cerebral boxes.

The first chapters of the book build the case for the problem to be solved: "Our understanding of the role of a specific region needs to be gradually bootstrapped." Pessoa initially dives into brain anatomy because, no matter how dry, it is a must. In turn, proper anatomy calls for embryology. And, as tackled later in the book, evolution also informs brain organization. Disciplines, we learn, are entangled too.

So, what is the remedy for reductionism? Pessoa goes for large-scale distributed circuits within a network perspective—a complex systems approach where "many relatively simple interacting parts" exhibit "emergent" behaviors. Emergence can be invoked as a free miracle, and the misuse of networks lends itself to hairball graphs. However, Pessoa's amalgamation of systems theory, cybernetics, and network science is a necessary step.

Pessoa claims that "biology does not work like physics, and even less so like engineering." He challenges linchpin assumptions in the life and mind sciences—the reducibility of organisms and their brains, ceteris paribus, and the belief that truth is to be found in simplicity. The obviously nontrivial requires restating: We cannot explain "all biology in terms of physics and chemistry."

Pessoa devotes many pages to his own research on emotion. The reader is then introduced to the role of the hypothalamus, the amygdala, the insula, and other brain areas such as the cingulate and prefrontal cortex. Wondering about neural modularity, Pessoa discusses the logic of dissociations. One can learn tremendously from lesions and manipulations, and yet brain regions can carry out tasks previously performed by other parts now gone.

Pessoa then challenges the "*billiard ball* model of causation" with a more dynamic view. However, he wonders whether neural trajectories, as signatures of cognitive tasks, do explanatory or descriptive work. He subscribes to the current renaissance of process philosophy in biology, whereby organisms are conceived of as processes rather than things.

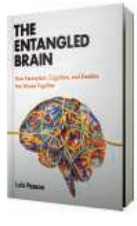
Much like classical physicists chiseling atoms, the irrepressible desire of neuroscientists to literally observe the mind leads to paradox, if not fallacy. Visualization and naming engross us. The fanfare provoked by a brain region lighting up in blue under a scan upon the presentation of a blue stimulus borders on a Monty Python sketch. And the lesson goes beyond brain areas: There is no such thing as the "gene of jealousy" or the "hormone of hate," for example.

Despite explicitly mentioning the pioneering work of theoretical biologists Ludwig von Bertalanffy and Robert May, Pessoa hints at the sterility of "idle armchair musing." Should we tend toward frenzied benchwork productions instead? The former without the latter is barren; the latter devoid of the former is bovine.

The book says little about the bodies, minds, and reciprocal interactions between organisms and their environments. Certainly, "a brain can be thought of as an entire circuit 'in between' sensory and motor cells." Perception, nevertheless, is virtual action.

The Entangled Brain often reads more like a manifesto than an argument. Indeed, mantras such as "circuit X produces behavior Y" betray a can of misconceptions. However, there is room for both separation and connection, as the synapse metaphorically symbolizes and literally enables. The disputes between lumpers and splitters are half-truths.

Given Pessoa's wink at a processual view of life, one wonders whether his post-reductionism also calls for a postmaterialist neuroscience. Paraphrasing Erwin Schrödinger, if *verschränkung* (entanglement) is the defining characteristic of brains and minds, it enforces an entire departure from classical lines of thought. In this sense, *The Entangled Brain* instantiates yet another conservative revolution in current neuroscience. ■



The Entangled Brain
Luiz Pessoa
MIT Press, 2022.
280 pp.

The reviewer is at the Instituto de Neurociencias, Consejo Superior de Investigaciones Científicas–Universidad Miguel Hernández de Elche, Alicante, Spain. Email: agomezmarin@gmail.com

10.1126/science.ade8689

PHOTO: SIMON RITZMANN/GETTY IMAGES

GENETICS

Genetic engineering's contested ethics

Good intentions at the intersection of principles, policy, and profit make for a bumpy road

By **Luis A. Campos**

In the fall of 1972, heralded by a flurry of new research, "the age of genetic engineering had begun, and no one seemed to care," observes Matthew Cobb in *As Gods: A Moral History of the Genetic Age*, as he describes the advent and rise of recombinant DNA technologies, the presumed fulfillment of the dreams of decades past. But everyone began to care, soon enough.

The 1975 Asilomar meeting on potential biohazards serves as a recurring touchstone for Cobb's remarkable jaunt through the twists and turns of the genetic engineering revolution. "Organising an international conference is a pain in the backside at the best of times," he notes. "Organising an international conference with the world's press clamouring to attend must be hell." The book is clearly not your father's foray through a fog of subtle distinctions and philosophical jargon.



Maxine Singer, Norton Zinder, Sydney Brenner, and Paul Berg converse at the 1975 Asilomar conference.

Driven by three "areas of concern"—heritable human gene editing, gene drives, and pathogen manipulation—*As Gods* ranges from recombinant DNA to synthetic biology and from "Frankenfood" and biological weapons to gene therapy and #CRISPRbabies. But local color is also abundant. We learn, for example, about the development of a "steampunk-sounding gene gun" at Stanford University in 1987 ("the laboratory soon smelled like a mixture of a McDonald's and a firing range"), as well as Barbra Streisand's genetically identical dogs and Greenpeace's foiled efforts to kidnap Dolly, the cloned sheep.

As a drosophilist with a journalist's gimlet eye for uncovering both technical subtleties and situational irony, Cobb interweaves tales of practicalities and moralities, with as careful attention to scientific achievements as to regulatory and social contexts. He is keenly aware of the happenstance role of money and politics, and he also has a knack for telling descriptions: Genentech in its earliest days is described as "barely more than a rented office and some headed notepaper," while Monsanto's many and repeated blunders connect to its "reputation for unthinking ambition."

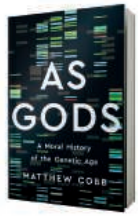
Cobb astutely captures sudden transformations, as with the advent of gene patenting, when "what had long been seen as either immoral or unconstitutional became widely accepted. Life itself could be owned, and everyone thought it was normal." And he observes how even well-meaning attempts at biosecurity related to bioweapons research may lead inadvertently to the opposite. Through its various case studies, *As Gods* shows that the ethical minefield is not some rhetorical flourish—it is the very world of everyday genetic engineering research.

"When scientists muse about the political and sociological consequences of their work, the results are often unedifying," Cobb complains. His moral elbows are sharp, and he is quick to point out errors and hypocrisies, whether from biotech capitalists' unholy mixtures of "overexcitement and ill-considered hyperbole," journalists who have "little grasp of either ethics or genetics," or pious musings from lab workers and ethicists alike. It is through careful attention to the details of complex circumstances that the book achieves clarity from what would otherwise be a muddled morass of detailed technical issues, ethical jargon, and competing interests.

Cobb also calls for respect and engagement with voices, views, and concerns coming from beyond laboratory walls or bureaucratic beltways. He notes that some fears, for instance those of genetically modified crops, "are apparently rooted deeper than facts can reach," even as he concludes that "neither utopia nor catastrophe has materialised."

While Cobb constantly grapples with the shifting boundaries of the technical and the ethical and is not always consistent in his conclusions—it is less a moral history than one that engagingly moralizes—his renderings of the relations, entanglements, and historical ironies are told with a keen eye to scientific details, finances, and contexts that matter. He shows time and again how complicated social dynamics of science cannot always be easily solved, even if matters often look clearer in retrospect: "Real-world problems, most of them social and not amenable to simple technofixes, have repeatedly brought a sharp dose of reality to the dreams of the genetic engineers," he notes midway through the book, before concluding again near the end: "We need to temper the visions of utopia, as well the fears of future dystopia."

Deftly navigating between the Scylla of inscrutable pronouncements from policymakers on task forces and the Charybdis of banal sureties from interested practitioners, Cobb's integration and synthesis of the work of many other scholars in a variety of fields who have explored these issues within science, within journalism, and within the worlds of scholarship make *As Gods* a valuable new go-to source that avoids common pieties and ritual invocations. ■



As Gods
Matthew Cobb
Basic Books, 2022.
464 pp.

The reviewer is Baker College Chair for the History of Science, Technology, and Innovation in the Department of History, Rice University, Houston, TX 77005, USA.
Email: lc@rice.edu

10.1126/science.ade5848

PHOTO: ALBUM/ALAMY STOCK PHOTO



The southern greater glider (*Petauroides volans*) population's habitat could be destroyed by logging.

Edited by **Jennifer Sills**

Extinction risk for Australia's iconic glider

Endemic to the eucalypt forests along the east coast of Australia, the southern greater glider (*Petauroides volans*) is rapidly declining due to ongoing land clearing, logging, and anthropogenic-driven climate change events, including the megafires in 2019 and 2020 (*1*). In May 2016, greater gliders were federally listed as Vulnerable to extinction (*2*) because the protections imposed to reduce the impacts of climate change and logging were not adequate to ensure the species' recovery (*3*). However, in the 6 years since the listing, there have been no changes to protect the greater glider. In 2022, the greater glider was uplisted to Endangered (*4*). To protect this species, we must prevent further logging of Australia's native forests.

Greater glider population locations (*5*) overlap considerably with logging that has been approved to occur between 2022 and 2026 (*6*). Logging in greater glider habitats will hasten the species' alarming downward trajectory (*7*). Amid the sixth mass extinction (*8*), it is imperative that we protect threatened species by applying a precautionary approach when considering further destruction to their habitats. Despite evidence that Australia's biodiversity is suffering major declines, out-of-date policies still reflect the exploitative paradigm upon which several of the country's industries have been built. Australian forests are not an infinite resource; once they are logged, they often take several centuries to recover (*9*).

In September, Australia pledged to reverse biodiversity loss by 2030 (*10*), including efforts to reduce deforestation. It is now up to Australians to choose whether to continue down the extinction path for the greater glider and many other Australian species or to protect our precious forests from further destruction.

Kita Ashman[1,2]* and Michelle Ward[1,3]
[1]World Wide Fund for Nature Australia, Melbourne, VIC 3000, Australia. [2]The Gulbali Institute, Charles Sturt University, Albury, NSW 2640, Australia. [3]Centre for Biodiversity and Conservation Science, The University of Queensland, St. Lucia, QLD 4072, Australia.
*Corresponding author. Email: kashman@wwf.org.au

REFERENCES AND NOTES

1. S. Legge *et al.*, *Glob. Ecol. Biogeogr.* **31**, 2085 (2022).
2. C. McLean *et al.* *For. Ecol. Manag.* **415–416**, 19 (2018).
3. K. R. Ashman, D. J. Watchorn, D. B. Lindenmayer, M. F. J. Taylor, *Pacific Conserv. Biol.* **28**, 277 (2022).
4. Department of Climate Change, Energy, Environment and Water, *Petauroides volans*—Greater Glider (southern and central) Glossary SPRAT Profile (Australian Government, Canberra, 2022); https://www.environment.gov.au/cgi-bin/sprat/public/publicspecies.pl?taxon_id=254.
5. Victorian Government, Forest Information Portal (2022); https://maps.ffm.vic.gov.au/fip/index.html?viewer=fip. On the left, click "Find a species" and search for Southern Greater Glider.
6. VicForests, Timber Release Plan (2022); https://www.vicforests.com.au/timber-release-plan.
7. D. B. Lindenmayer *et al.*, *Biol. Conserv.* **144**, 1663 (2011).
8. G. Ceballos, P. R. Ehrlich, *Science* **360**, 1080 (2018).
9. E. J. Bowd, S. C. Banks, C. L. Strong, D. B. Lindenmayer, *Nat. Geosci.* **12**, 113 (2019).
10. Leaders pledge for nature (2022); https://www.leaderspledgefornature.org/.

10.1126/science.adf1013

Boost Egypt's coral reef conservation efforts

The Intergovernmental Panel on Climate Change (IPCC) predicts that 70 to 90% of warm-water reefs will disappear this century even if warming is constrained to 1.5°C (*1*). But the corals of the northern Red Sea are thermally resilient and likely to survive IPCC warming projections (*2*, *3*). Egypt's territorial waters contain about 1800 km of fringing reef (*4*) and include almost the entire western half of the resilient region in the northern Red Sea. Although Egypt's corals can tolerate the rising temperatures that are decimating reefs elsewhere, they face severe local threats, including unsustainable tourism, coastal development, sewage discharge, and desalination plant discharge (*5*). Interventions are urgently needed to improve coral conservation.

Fringing reefs are of high cultural and economic importance. Egypt has the most valuable coral reef tourism economy in the world, contributing 2% of its GDP (*6*). The reefs could benefit from an expanded and fortified marine protected area network, which currently protects only 4% of its waters (*7*). Fisheries management and enforcement in the country are inadequate as well (*8*). Egypt should also prioritize investment in sustainable tourism practices and infrastructure that mitigates land-based pollution, such as wastewater treatment infrastructure and garbage disposal mechanisms.

International nongovernmental organizations (NGOs) and multipartner initiatives have facilitated finance mechanisms for reef conservation in many regions (*9*, *10*), and Egypt could benefit from similar partnerships. However, there remains scant international engagement in Egypt compared with other coral reef countries. The United Nation's 27th Climate Change Conference of the Parties, which has convened on Egypt's Red Sea coast, presents an excellent opportunity to boost the country's reef conservation efforts. Egypt's

PHOTO: © JOSH BOWELL

government should leverage this event to forge new collaborations between governments, research institutions, NGOs, and local communities. Together, this coalition can advance a shared commitment to conserve one of the few coral reef refuges from climate change.

Karine Kleinhaus[1]*, John J. Bohorquez[1,2], Yasser M. Awadallah[3], David Meyers[2], Ellen Pikitch[1]
[1]Stony Brook University, School of Marine and Atmospheric Sciences, Stony Brook, NY 11790, USA. [2]Conservation Finance Alliance, Bronx, NY 10460, USA. [3]Egyptian Environmental Affairs Agency, Sharm El Sheikh, South Sinai Governorate 46, Egypt.
*Corresponding author.
Email: karine.kleinhaus@stonybrook.edu

REFERENCES AND NOTES

1. IPCC, in Global Warming of 1.5°C: An IPCC Special Report on the Impacts of Global Warming of 1.5°C above Pre-Industrial Levels and Related Global Greenhouse Gas Emission Pathways, in the Context of Strengthening the Global Response to the Threat of Climate Change, Sustainable Development, and Efforts to Eradicate Poverty, V. Masson-Delmotte *et al.*, Eds. (Cambridge University Press, 2018), pp. 3–24.
2. M. Fine, H. Gildor, A. Genin, *Glob. Change Biol.* **19**, 3640 (2013).
3. E. O. Osman *et al.*, *Glob. Change Biol.* **24**, e474 (2018).
4. M. Khaled, *Int. J. Engin. Educ.* **4**, 17 (2019).
5. A. D. Shepherd, A. S. M. Khalil, M. A. Amer, "State of the marine environment report for the Red Sea and Gulf of Aden (SOMERSGA II)" (The Regional Organization for the Conservation of the Environment of the Red Sea and Gulf of Aden, 2020); http://persga.org/Documents/Publications/QR_Downloads/English/SOMERSGA_2020.pdf.
6. M. Spalding *et al.*, *Mar. Pol.* **82**, 104 (2017).
7. Marine Protection Atlas, Marine Protection by Country/Egypt (2022); https://mpatlas.org/countries/EGY.
8. M. Samy-Kamal, *Rev. Fish Biol. Fish.* **25**, 631 (2015).
9. J. J. Silver, L. M. Campbell, *Int. Soc. Sci. J.* **68**, 241 (2018).
10. R. Victurine *et al.*, "Conservation finance for coral reefs: A vibrant oceans initiative whitepaper" (Wildlife Conservation Society, Bronx, NY, 2022); http://wcs.org/coral-finance-whitepaper.

COMPETING INTERESTS

K.K. is president of the Red Sea Reef Foundation. J.J.B. and D.M. are paid consultants for the Global Fund for Coral Reefs.

10.1126/science.adf3377

China must balance renewable energy sites

China leads global renewables installation (*1*, *2*). In 2021, China's solar and wind installed capacity was 306.4 GW and 329 GW, respectively, accounting for 36.3% and 39.9% of the global market (*3*). However, enthusiasm for installed capacity obscures insufficient penetration into some areas of the country, which hinders the potential benefits of wind and solar energy. In 2021, the average waste rates of China's wind and solar were 3.1% and 2.0%, respectively, and more than 10% in Qinghai Province (*4*).

Geographic imbalance is a challenge to China's onshore wind and photovoltaic market. The richest wind and solar resources are located in the northern and western regions of China, far from the high-demand, population-dense areas (*5*). Provinces in the coastal region consume approximately 80% of total national electricity (*6*), which is problematic given that the vast majority of the installed capacity is separated by a distance of more than 1000 miles. Long-distance transmission is highly inefficient; an ultrahigh-voltage line loses 3.5% of power or more as it travels (*6*, *7*). As planned, China is building massive wind and solar power bases in the Gobi and other desert areas. Unfortunately, energy storage facilities and transmission channels are difficult to achieve in these areas, and without them, even more renewable energy is likely to be wasted (*8*).

The central government encourages local governments to build renewables but sets penetration targets for each province to avoid unchecked construction. As a result, some provinces have suspended wind power and photovoltaic projects, especially projects that cannot be completed because of land use or COVID-19 policies. By 13 September, 120 projects with a total scale of 6492 MW had been abolished in Hebei Province, Shanxi Province, and Shaanxi Province, representing 2267.5 MW and 4224.5 MW of wind and photovoltaic projects, respectively (*9*).

To fundamentally improve renewable energy penetration, China must prioritize energy storage technologies such as pumped storage hydropower and virtual synchronous machine technology (*10*, *11*), which will allow the infrastructure currently in development to provide power to distant regions. The country also needs to build transregional high-capacity transmission channels and flexible grids (*12*). Meanwhile, the government should seek new market solutions and break down the barriers of China's special regional power market by establishing an interprovincial power trading market mechanism (*8*). Finally, it is necessary to strengthen the development of distributed renewables and offshore wind power in the east (*6*) to alleviate the mismatch between renewable energy construction and market demand.

Yu Yang[1,2,3]* and Siyou Xia[1,2]
[1]Institute of Geographic Science and Natural Resources Research, Chinese Academy of Sciences, Beijing, China. [2]College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, China. [3]Institute of Strategy Research of Guangdong-Hong Kong-Macao Greater Bay Area, Guangzhou, China.
*Corresponding author.
Email: yangyu@igsnrr.ac.cn

REFERENCES AND NOTES

1. J. I. Lewis *et al.*, *Science* **350**, 1034 (2015).
2. C. Wang, F. Wang, *Science* **357**, 764 (2017).
3. BP, Statistical Review of World Energy (2022); https://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy.html.
4. National Energy Administration, "National renewable energy power development monitoring and evaluation report" (2021); www.nea.gov.cn/2022-09/16/c_1310663387.htm [in Chinese].
5. C. Wang *et al.*, *Renew. Sust. Energ. Rev* **134**, 110337 (2020).
6. P. Sherman *et al.*, *Sci. Adv.* **6**, eaax9571 (2020).
7. S. Zhang, *Nature* **514**, 168 (2014).
8. J. Liu *et al.*, *Engineering* **7**, 1611 (2021).
9. Polaris Solar PV, "120 stock wind and photovoltaic projects (6.5GW) in three provinces were abolished (List Attached)" (2022); https://guangfu.bjx.com.cn/news/20220913/1254435.shtml [in Chinese].
10. S. O'Meara, Y. Ye, *Nature* **603**, S41 (2022).
11. X. Zheng, "Pumped storage hydropower to bloom in China," *China Daily* (2021); www.chinadaily.com.cn/a/202109/29/WS6153f4d8a310cdd39bc6c5b9.html.
12. A. Alirezazadeh *et al.*, *Energy* **191**, 116438 (2020).

10.1126/science.adf3720

TECHNICAL COMMENT ABSTRACTS

Comment on "Models predict planned phosphorus load reduction will make Lake Erie more toxic"

Jef Huisman, Elke Dittmann, Jutta Fastner, J. Merijn Schuurmans, J. Thad Scott, Dedmer B. Van de Waal, Petra M. Visser, Martin Welker, Ingrid Chorus

Hellweger *et al*. (Reports, 27 May 2022, pp. 1001) predict that phosphorus limitation will increase concentrations of cyanobacterial toxins in lakes. However, several molecular, physiological, and ecological mechanisms assumed in their models are poorly supported or contradicted by other studies. We conclude that their take-home message that phosphorus load reduction will make Lake Erie more toxic is seriously flawed.

Full text: dx.doi.org/10.1126/science.add9959

Response to Comment on "Models predict planned phosphorus load reduction will make Lake Erie more toxic"

Ferdi L. Hellweger, Charlotte Schampera, Robbie M. Martin, Falk Eigemann, Derek J. Smith, Gregory J. Dick, Steven W. Wilhelm

Huisman *et al*. claim that our model is poorly supported or contradicted by other studies and the predictions are "seriously flawed." We show their criticism is based on an incomplete selection of evidence, misinterpretation of data, or does not actually refute the model. Like all ecosystem models, our model has simplifications and uncertainties, but it is better than existing approaches that ignore biology and do not predict toxin concentration.

Full text: dx.doi.org/10.1126/science.ade2277

RESEARCH

IN *SCIENCE* JOURNALS

Edited by **Michael Funk**



False-color image of a fossilized tree fern showing the inner xylem network (blue)

PHOTO: LUDWIG LUTHARDT, MUSEUM FUR NATURKUNDE, BERLIN

PLANT MORPHOLOGY

Drought shapes plant architecture

Since plants colonized land, they have developed increasingly complex vessel architectures to carry water from their roots to their highest leaves. Vascular plants now display a diversity of xylem strand shapes in cross section, from elliptical to linear to many lobed. Bouda *et al.* investigated whether selection from drought, which causes vessel cavitation and embolism, drove the complexity of xylem strand shape as plants inhabited drier climates. By simulating embolism spread between vessels across varying shape and complexity, including those seen in extant lycophytes and ferns and extinct plant fossils, the authors found that evolutionary changes in xylem strand shape have reduced embolism spread and made plants less vulnerable to drought. —BEL

Science, add2910, this issue p. 642

DRUG DEVELOPMENT

An oral route to TNF inhibition

Excessive production of the proinflammatory cytokine tumor necrosis factor (TNF) drives many inflammatory diseases. Current targeted therapies for these conditions consist mainly of costly biologics that must be injected. Javaid *et al*. identified a nontoxic small-molecule inhibitor called TIM1 that potently inhibited TNF signaling in mouse and human cells. Oral delivery of TIM1 or a more potent derivative improved symptoms and delayed disease progression in a mouse model of inflammatory arthritis to a similar extent as injection of the US Food & Drug Admistration–approved TNF-targeting biologic etanercept. —AMV

Sci. Signal. **15**, eabi8713 (2022).

RAINFALL EXTREMES

A hard rain is falling

Short-duration, extreme rainfall can cause dangerous flash flooding, threatening life, infrastructure, and the landscape. Studies of this type of event have focused mainly on daily rain totals, not considering how precipitation might vary on shorter time scales. Ayat *et al*. analyzed subhourly rainfall extremes near Sydney, Australia, over 20 years and found that they are increasing much faster than those over longer periods. Better understanding of such extremes is vital for effective climate adaptation and to reduce the vulnerability of populated regions. —HJS

Science, abn8657, this issue p. 655

CANCER

Cell immortality gets a boost

Telomeres are DNA sequences that cap the ends of chromosomes and become shorter as cells divide. The enzyme telomerase maintains telomere length so that cells can continue dividing. Cancer cells often have high telomerase activity, and noncoding mutations in the *TERT* gene (which encodes telomerase) are frequently found in tumors. Chun-on *et al*. studied melanomas and identified mutations in the promoter of *TPP1*, which encodes the telomere-binding protein TPP1 that recruits telomerase to the telomere. Such promoter mutations created a transcription factor site similar to mutations previously identified in the *TERT* gene promoter. Co-expression of TERT and TPP1 leads to synergistic telomere lengthening, indicating that *TPP1* and *TERT*

promoter mutations cooperate to immortalize melanoma cells. —PNK

Science, abq0607, this issue p. 664

ASTROPHYSICS

Stars constrain the fine-structure constant

The strength of the electromagnetic force is quantified by the fine-structure constant α. The Standard Model of particle physics provides no explanation for its value, which could conceivably vary from place to place. Murphy *et al*. have used spectra of 17 nearby stars, with properties matched to the Sun, to investigate absorption lines that are sensitive to α. They set an upper limit of 50 parts per billion on variations of α between the stars. The results rule out substantial changes in α within the local region of the Milky Way, filling a gap between laboratory measurements and constraints from the distant Universe. —KTS

Science, abi9232, this issue p. 634

CORONAVIRUS

Defending against Omicron

The Omicron BA.1 lineage of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) emerged in late 2021 and quickly became dominant, in part because of a large number of mutations that allowed escape from existing antibodies. New infection waves have come from other Omicron sublineages. Park *et al*. found that either a vaccination booster or a breakthrough infection elicits neutralization activity against the Omicron variants, but only a breakthrough infection induces an antibody response in the nasal mucosa, which might give better protection against transmission. Testing a panel of antibodies, the authors showed that the antibody S2X324 potently neutralizes all SARS-CoV-2 variants tested, making it a candidate for therapeutic development. A cryo–electron microscopy structure shows how this antibody accommodates Omicron-specific mutations to block binding of the viral spike protein to the human ACE2 receptor across the variants. —VV

Science, adc9127, this issue p. 619

METALLURGY

Getting rid of the creep

Materials can plastically deform by creep, which is amplified at higher temperatures. Avoiding creep often requires making large single crystals of an alloy, which is expensive and time consuming. Zhang *et al*. show that introducing astable grain boundary network into a nanograined medium-entropy alloy also improves the creep behavior at high temperature. The resulting alloy has high creep resistance even under high stresses, a important property in structural alloys. —BG

Science, abq7739, this issue p. 659

STRUCTURAL BIOLOGY

A group II intron ready to attack

By forming ribonucleoprotein (RNP) complexes with specialized reverse transcriptases, group II introns can splice out of RNA and insert themselves into new DNA sites. Chung *et al*. used cryo–electron microscopy to investigate how an ancient class of group II intron retroelements recognize the shape and sequence of a highly structured DNA target, thereby revealing new molecular recognition strategies between RNPs and DNA. Structural comparison of the isolated RNP with that of the DNA-bound holoenzyme reveals that the group II intron RNP is primed to attack its DNA target without major conformational rearrangements. The study sheds light on retroelement structure, function, and proliferation. —DJ

Science, abq2844, this issue p. 627

IN OTHER JOURNALS

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



PHOTO: NASA

CLOUDS

The big picture

Cloudiness is a fundamental determinant of Earth's energy balance. Shortcomings in our ability to faithfully represent cloud coverage on a global scale are therefore a major source of uncertainty in climate models. Datseris *et al.* developed a conceptual model of global cloudiness that uses only a few basic state variables such as surface temperature and pressure velocity to calculate the spatiotemporal distribution of clouds over the whole planet. In addition to helping to develop a better theoretical understanding of clouds, their work should provide insight into how cloudiness will change in a changing climate. —HJS

Geophys. Res. Lett. 10.1029/2022GL099678 (2022).

Global cloudiness can be approximated from a small array of basic atmospheric parameters.

ELECTRON MICROSCOPY

Focusing an electron beam with light

Like everyday photography, it is possible to increase the brightness of an electron beam or prolong the exposure time, thus improving the picture quality in electron microscopy. However, for living biological tissues, too intense a beam could harm the sample, and for experiments in which high temporal resolution is needed, a longer exposure time would mean a lower frame rate. Mihaila *et al.* used a laser beam to shape an electron beam to improve its focus. Their setup uses a spatial light modulator to control the cross-sectional pattern of a laser beam, which crosses paths with an electron beam to shape its electron distribution. The modulator provides a means to program and pattern electron beams with submillimeter resolution, which can help to lower the electron exposure needed to image a living biological sample. —YY

Phys. Rev. X **12**, 031043 (2022).

THERAPEUTICS

Opening the gates for antibodies

Delivery of therapeutics into the central nervous system is notoriously difficult because of

ALSO IN *SCIENCE* JOURNALS

Edited by **Michael Funk**

MICROBIOLOGY

Rewilding microbiota in crops

As plants have been domesticated and bred to produce high-performance crops, they have become less able to establish symbiotic relationships with microbes that can promote nutrition and resistance to stress. In a Perspective, Raaijmakers and Kiers discuss the emerging idea that restoring microbial interactions observed in the wild plants from which our modern crops are derived could improve their health and sustainability. The authors discuss the features in plants that allow symbiotic relationships with microbes to be established and the benefits of examining these mechanisms in wild plants in their native environments. They also propose a path forward to allow the development of "microbiome-assisted" crops. —GKA

Science, abn6350, this issue p. 599

RADIO ASTRONOMY

Fast radio bursts

In 2007, astronomers studying archival data serendipitously identified a bright flash of radio waves, lasting a few milliseconds, that was apparently of extragalactic origin. This previously unknown type of signal is now called a fast radio burst (FRB). Bailes reviews the discovery of FRBs and the subsequent rapid expansion in our understanding of them. More than 800 FRB sources have now been observed, some of which are known to repeat. Several FRBs (repeating and nonrepeating) have been located to their host galaxies, showing that they originate from a variety of localities. Multiple lines of evidence indicate that FRBs are probably emitted by magnetized neutron stars, although the physical mechanism is still unclear. —KTS

Science, abj3043, this issue p. 615

IMMUNOLOGY

A Mediator for T cell function

T lymphocytes are white blood cells that have the ability to fight cancer. Potent T cell responses are the cornerstone of successful cancer immunotherapy; however, T cells can become worn out over time and lose their ability to attack cancer. Searching for ways to improve T cell immunotherapy, Freitas *et al*. performed a genome-wide CRISPR screen of human chimeric antigen receptor (CAR) T cells (see the Perspective by Zebley and Youngblood). Mediator complex subunit 12 (MED12) and cyclin C (CCNC), components of the Mediator cyclin-dependent kinase module, were found to be key regulators of T cell activation and effector function. When MED12 or CCNC was genetically inactivated in CAR T cells, increased T cell expansion, metabolic fitness, and tumor control were observed. —PNK

Science, abn5647, this issue p. 616; see also adf0546, p. 598

FISHERIES

Conservation works

Tuna and billfishes are large species that have long been targeted by fisheries, whereas sharks, which are also large fishes, have tended to be considered as by-catch or nontarget species. Juan-Jorda *et al*. used an approach that monitors yearly changes in the International Union for Conservation of Nature Red List status to estimate population status for these three groups (see the Perspective by Burgess and Becker). After almost three decades of decline, tuna and billfishes have begun to recover because of proactive fisheries management approaches. Sharks, however, which have received much less conservation attention, have continued to decline. These results both reinforce the value of conservation and management and emphasize the need for immediate implementation of these approaches for sharks. —SNV

Science, abj0211, this issue p. 617; see also add0342, p. 596

AIR POLLUTION

Powering up

India is the world's third-largest producer of carbon dioxide even though its per capita emissions are very low. As India's population becomes more affluent and consumes more energy, their power sector, now heavily reliant on coal, will grow. Therefore, greenhouse gas emissions and other types of air pollution likely will also grow unless large changes are made to the electricity production sector. Sengupta *et al*. present a highly resolved model of Indian power generation and demand that assesses the emission impacts of various power sector policy interventions in India (see the Perspective by Deshmukh and Chatterjee). This analysis provides valuable guidance about the development of the power sector and the costs associated with different development pathways. —HJS

Science, abh1484, this issue p. 618; see also ade6040, p. 595

FLEXIBLE ELECTRONICS

Acoustic patterning and fabrication

Liquid metals can be used to form the conductive pathways in a flexible matrix, but this approach requires patterning of the soft material and sintering of the liquid metal using lasers or mechanical force. Lee *et al*. used acoustic fields to assemble a network of liquid metal particles inside a polymer matrix for the fabrication of elastic printed circuit boards (see the Perspective by Qiao and Tang). Their devices showed high conductivity, high stretchability, strong adhesiveness, and negligibly small changes in electrical resistance during stretching. Because the acoustic field strategy is universal, the authors synthesized hydrogels, a self-healing elastomer, and photoresists by combining various polymers with liquid metals. —MSL

Science, abo6631, this issue p. 637; see also ade1813, p. 594

NEUTRON STARS

Polarization constrains magnetar emission

Magnetars are young neutron stars with high magnetic fields that are usually observed at x-ray wavelengths. The emission mechanism and geometry of the emitting region have been unclear. Taverna *et al*. measured the x-ray polarization of the magnetar 4U 0142+61. The polarization degree and angle change as a function of x-ray energy, indicating two different emission regions. The authors preferred a model in which most of the x-rays are emitted by an equatorial band on the surface of the neutron star, with some of the photons then being scattered to higher energies by collisions with electrons in the surrounding magnetic field. —KTS

Science, add0080, this issue p. 646

BLACK HOLES

X-ray polarization of Cygnus X-1

A black hole in a binary system can rip material off of its companion star, which heats up and forms an accretion disk. The disc emits light in the optical and x-ray bands, forming an x-ray binary (XRB) system. Some XRBs also launch a jet of fast-moving material that is visible at radio wavelengths. Krawczynski *et al*. observed the x-ray polarization of Cygnus X-1, a black hole XRB with a radio jet. By comparing the measured polarization properties with several

competing XRB models, they eliminated some hypothesized geometries and determined that the x-ray–emitting region extends parallel to the accretion disc. —KTS

Science, add5399, this issue p. 650

INFLAMMATION

NLRP1's danger-sensing mechanism

Inflammasomes are multiprotein cytoplasmic complexes that sense danger signals. NLRP1 inflammasomes can be experimentally activated by several stimuli, but the core endogenous danger signal triggering NLPR1 activation has remained elusive. Ball *et al.* used a proteomics approach to identify proteins bound to the N-terminal regulatory regions of human NLRP1 and discovered that activation is normally suppressed by binding of the oxidized form, but not the reduced form, of the thioredoxin-1 protein. Under cellular conditions in which reactive oxygen species are in short supply, also known as reductive stress, depletion of oxidized thioredoxin-1 causes increased NLRP1 activation. These findings provide new insights into the full range of cellular parameters that the innate immune system can sense as it surveys the intracellular environment for unwelcome danger signals. —IRW

Sci. Immunol. **7**, eabm7200 (2022).

CANCER IMMUNOTHERAPY

Dynamic bispecific duo

Bispecific antibodies are being used for cancer immunotherapy because they target two different antigens, thus facilitating T cell targeting to tumors. Odronextamab is a CD20xCD3-bispecific antibody that shows promising activity in patients with diffuse large B-cell lymphoma (DLBCL), but not all patients achieve complete responses and there remains a high unmet need in the setting of relapsed or refractory disease. Wei *et al*. demonstrate that REGN5837, a CD22xCD28-bispecific antibody, enhanced the antitumor activity of odronextamab in preclinical models of DLBCL, and that the combination exhibited no toxicity in primates while augmenting T cell activation. $CD28^+CD8^+$ T cells were expanded in non-Hodgkin lymphoma samples from a phase 1 odronextamab trial, suggesting that the addition of REGN5837 could enhance antitumor activity. These findings highlight the potential clinical usefulness of a bispecific antibody combination for treating DLBCL. —CNF

Sci. Transl. Med. **14**, eabn1082 (2022).

promoter mutations cooperate to immortalize melanoma cells. —PNK

Science, abq0607, this issue p. 664

ASTROPHYSICS

Stars constrain the fine-structure constant

The strength of the electromagnetic force is quantified by the fine-structure constant α. The Standard Model of particle physics provides no explanation for its value, which could conceivably vary from place to place. Murphy *et al*. have used spectra of 17 nearby stars, with properties matched to the Sun, to investigate absorption lines that are sensitive to α. They set an upper limit of 50 parts per billion on variations of α between the stars. The results rule out substantial changes in α within the local region of the Milky Way, filling a gap between laboratory measurements and constraints from the distant Universe. —KTS

Science, abi9232, this issue p. 634

CORONAVIRUS

Defending against Omicron

The Omicron BA.1 lineage of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) emerged in late 2021 and quickly became dominant, in part because of a large number of mutations that allowed escape from existing antibodies. New infection waves have come from other Omicron sublineages. Park *et al*. found that either a vaccination booster or a breakthrough infection elicits neutralization activity against the Omicron variants, but only a breakthrough infection induces an antibody response in the nasal mucosa, which might give better protection against transmission. Testing a panel of antibodies, the authors showed that the antibody S2X324 potently neutralizes all SARS-CoV-2 variants tested, making it a candidate for therapeutic development. A cryo–electron microscopy structure shows how this antibody accommodates Omicron-specific mutations to block binding of the viral spike protein to the human ACE2 receptor across the variants. —VV

Science, adc9127, this issue p. 619

METALLURGY

Getting rid of the creep

Materials can plastically deform by creep, which is amplified at higher temperatures. Avoiding creep often requires making large single crystals of an alloy, which is expensive and time consuming. Zhang *et al*. show that introducing astable grain boundary network into a nanograined medium-entropy alloy also improves the creep behavior at high temperature. The resulting alloy has high creep resistance even under high stresses, a important property in structural alloys. —BG

Science, abq7739, this issue p. 659

STRUCTURAL BIOLOGY

A group II intron ready to attack

By forming ribonucleoprotein (RNP) complexes with specialized reverse transcriptases, group II introns can splice out of RNA and insert themselves into new DNA sites. Chung *et al*. used cryo–electron microscopy to investigate how an ancient class of group II intron retroelements recognize the shape and sequence of a highly structured DNA target, thereby revealing new molecular recognition strategies between RNPs and DNA. Structural comparison of the isolated RNP with that of the DNA-bound holoenzyme reveals that the group II intron RNP is primed to attack its DNA target without major conformational rearrangements. The study sheds light on retroelement structure, function, and proliferation. —DJ

Science, abq2844, this issue p. 627

IN OTHER JOURNALS

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



PHOTO: NASA

CLOUDS

The big picture

Cloudiness is a fundamental determinant of Earth's energy balance. Shortcomings in our ability to faithfully represent cloud coverage on a global scale are therefore a major source of uncertainty in climate models. Datseris *et al.* developed a conceptual model of global cloudiness that uses only a few basic state variables such as surface temperature and pressure velocity to calculate the spatiotemporal distribution of clouds over the whole planet. In addition to helping to develop a better theoretical understanding of clouds, their work should provide insight into how cloudiness will change in a changing climate. —HJS

Geophys. Res. Lett. 10.1029/2022GL099678 (2022).

Global cloudiness can be approximated from a small array of basic atmospheric parameters.

ELECTRON MICROSCOPY

Focusing an electron beam with light

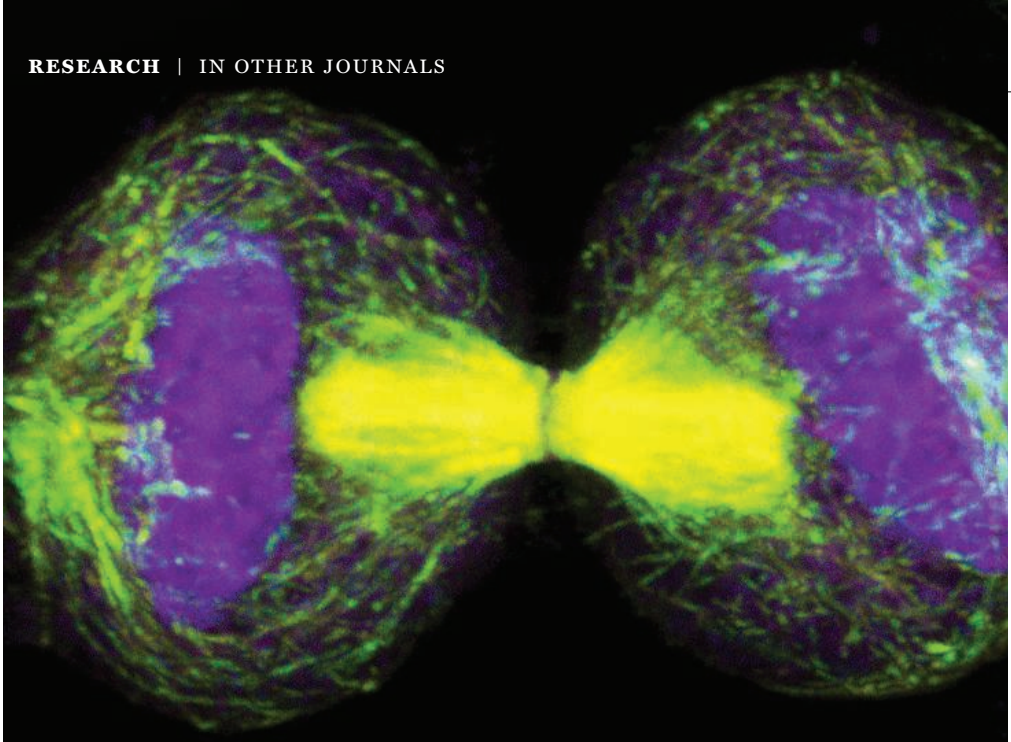
Like everyday photography, it is possible to increase the brightness of an electron beam or prolong the exposure time, thus improving the picture quality in electron microscopy. However, for living biological tissues, too intense a beam could harm the sample, and for experiments in which high temporal resolution is needed, a longer exposure time would mean a lower frame rate. Mihaila *et al.* used a laser beam to shape an electron beam to improve its focus. Their setup uses a spatial light modulator to control the cross-sectional pattern of a laser beam, which crosses paths with an electron beam to shape its electron distribution. The modulator provides a means to program and pattern electron beams with submillimeter resolution, which can help to lower the electron exposure needed to image a living biological sample. —YY

Phys. Rev. X **12**, 031043 (2022).

THERAPEUTICS

Opening the gates for antibodies

Delivery of therapeutics into the central nervous system is notoriously difficult because of



CELL BIOLOGY

Nuclear mechanosensing and cell division

Cell division depends on the biochemical activation and nuclear translocation of cyclin B1–CDK1 complexes. Dantas *et al.* used micromanipulation and high-resolution imaging to study the role of biomechanical forces as cultured mammalian cells approached mitosis. They found that during the transition from G_2 to mitosis, actomyosin contractility is transmitted to the nucleus through the Linker of Nucleoskeleton and Cytoskeleton (LINC) complex. This process triggers nuclear envelope unfolding and increased nuclear tension, which activates calcium-dependent phospholipase cPLA2 and results in faster translocation of cyclin B1 into the nucleus. This mechanical signal fine-tunes cyclin B1 transport across the nuclear pores, ensuring timely mitotic entry. Thus, the nucleus acts as a force sensor, regulating cell division according to the cellular tension state. —SMH *J. Cell Biol.* **221**, e202205051 (2022).

Mechanical cues from nuclear deformation trigger the biochemical events involved in the timing of cell division, shown in this light micrograph.

IMAGE: DR. GOPAL MURTI/SCIENCE SOURCE

the blood–brain barrier, which blocks most molecules from getting across. For antibody-based therapies, the challenges are even greater because they carry a high risk of triggering neuroinflammation after crossing the blood–brain barrier. Edavettal *et al*. addressed this issue by combining antibody-based therapeutics with transcytosis-enabling modules engineered to facilitate receptor-mediated transport into the brain. The authors tested their approach in multiple animal models and disease conditions, ranging from neurodegenerative disease to cancer metastasis, with promising results in each setting. —YN

Med (N Y). 10.1016/j.medj.2022.09.007 (2022).

PARTICLE PHYSICS

How wide is the Higgs?

The discovery of the Higgs boson 10 years ago provided a measurement of its mass. Measuring its lifetime, a property connected to the uncertainty in its mass through the Heisenberg principle, would enable a check against the Standard Model of particle physics. However, making a direct measurement is tricky because the uncertainty in the Higgs boson mass (its particle width) is predicted to be much smaller than the experimental resolution of the detectors at the Large Hadron Collider (LHC). Instead, the Compact Muon Solenoid collaboration at the LHC used comparative measurements of the off- and on-shell Higgs boson production to determine the boson's width and found that it agreed with the Standard Model prediction. —JS

Nat. Phys. 10.1038/s41567-022-01682-0 (2022).

INVASION ECOLOGY

Megaherbivores suppress invasion

Herbivores can play an important role in plant invasions by preferentially feeding on either the introduced or native species or by aiding in seed dispersal. Wells *et al*. investigated whether megaherbivores such as elephants facilitate or hinder the establishment of *Opuntia stricta*, a cactus that has spread prolifically across African savannas. Data from three long-term herbivore-exclusion experiments in Laikipia, Kenya, showed that *Opuntia* density was generally higher in sites with no megaherbivores. At the landscape scale, areas with higher herbivore diversity and elephant occurrence also had a lower probability of cactus occurrence. Megaherbivore feeding and disturbance (such as trampling) creates a net negative effect on cacti, contributing to biotic resistance to plant invasion. —BEL

J. Ecol. 10.1111/1365-2745.14010 (2022).

RACIAL DISPARITIES

Assisted reproductive disparities

In the United States, fetal and neonatal deaths are disproportionately higher among some racial and ethnic groups, particularly non-Hispanic Black women. Although the underlying causes of the disparities are not fully understood, it is assumed that differences in socioeconomic status contribute to it. However, little is known about racial and ethnic disparities among women who opt for medically assisted reproduction, which includes assisted reproduction technology such as in vitro fertilization. These procedures are largely used by economically privileged women because US health insurance companies rarely cover the costs. Lisonkova *et al*. examined this knowledge gap in a study of US singleton births between 2016 and 2017, when more than 90,000 infants were conceived by medically assisted reproduction. Analysis of births conceived by assisted reproduction technology revealed that, compared with infants of non-Hispanic White women, neonatal death was four times higher in infants of non-Hispanic Black women and nearly two times higher among Asian/Pacific Islander and Hispanic infants. This suggests that other unmeasured factors, including racism and institutional bias, may drive these disparities. —EEU

Pediatrics **150**, e2021055855 (2022).

RESEARCH ARTICLE SUMMARY

AIR POLLUTION

Subnational implications from climate and air pollution policies in India's electricity sector

Shayak Sengupta, Peter J. Adams, Thomas A. Deetjen, Puneet Kamboj, Swati D'Souza, Rahul Tongia, Inês M. L. Azevedo*

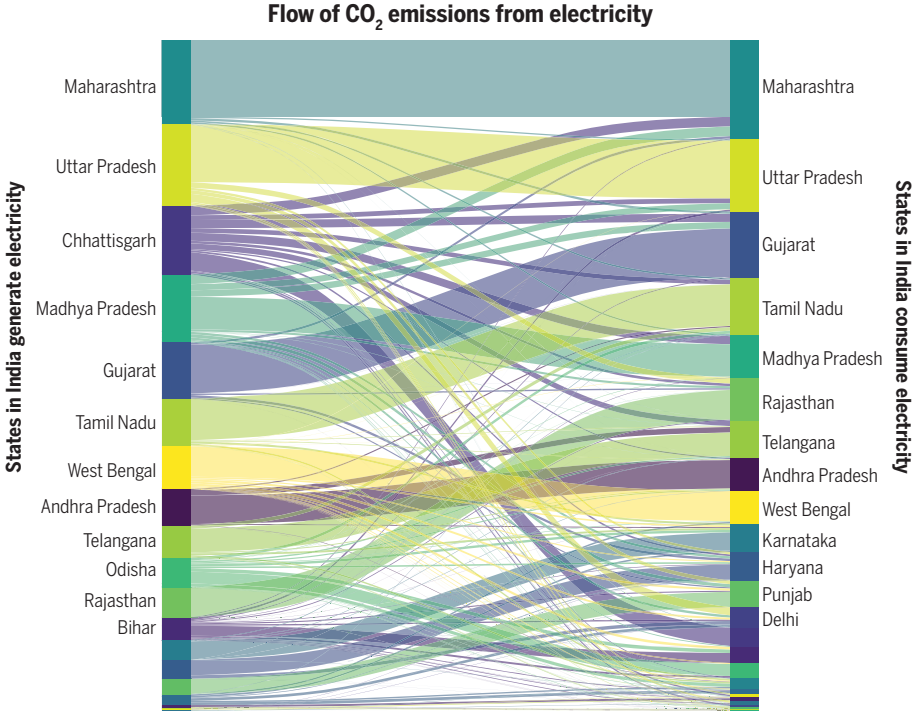
INTRODUCTION: India is the world's third-largest economy and power producer, with growing electricity demand from low–per capita electricity consumption. Despite the growth of renewable energy, coal-heavy electricity generation means that greenhouse gas and air pollutant emissions from the power sector are, and will remain, an important focus of public policies in India.

Each state in India largely schedules and dispatches its own power. Renewable electricity capacity is concentrated mostly in wealthier states in southern and western India. Meanwhile, hydroelectric power is located predominantly in Himalayan northern and northeastern India. Coal capacity is found throughout the country, but the cheapest coal plants are located near coal mines in a handful of poorer, eastern states. Electricity generation mixes vary by state, and both the central and state governments share jurisdiction over the power sector.

RATIONALE: A clear understanding of the emissions of greenhouse gases and air pollutants from the Indian grid at a subnational scale has not existed before now. No study has yet quantified the expected spatial heterogeneity arising from the current federal Indian power sector—i.e., which states are responsible for emissions based on the electricity that they generate or consume. We develop and present a reduced-order dispatch model of Indian power generation to assess CO_2 and SO_2 emission effects of policy interventions. The model simulates which power plants generate electricity to meet demand by minimizing short-term costs from power contracts in India. We focus on short-term effects for the current Indian grid and analyze how policies could induce spatial differences in emissions and costs between states in India.

RESULTS: We find that average nationwide CO_2 and SO_2 emissions intensities for electricity in India fail to capture the considerable heterogeneity between states. Electricity production and consumption and associated emissions between Indian states show orders-of-magnitude differences tied to population, akin to the differences between different countries around the world. A carbon tax results in little short-term emissions reductions because there is not enough dispatchable lower emission spare capacity to substitute coal. Moreover, it would disproportionately increase costs to poorer, coal-heavy eastern states. The implementation of sulfur controls will likely result in large reductions of SO_2 emissions, with the important outcome of reducing the current premature mortality associated with air pollution in India. Our simulations suggest that dispatching plants at the regional level rather than at the state level would lead to a small increase in both SO_2 and CO_2 emissions, with heterogeneous cost effects on states. Regionally coordinated dispatch would also impose changes in the spatial patterns of SO_2 and CO_2 emissions by shifting generation and emissions from distant plants to cheaper plants closer to eastern coal-mining regions.

CONCLUSION: Our analysis shows that policies that have modest or negligible emissions impacts at the aggregate, national level nonetheless have disparate state-level spatial emissions and cost effects. Electricity decarbonization and emissions reductions efforts in India must show an appreciation for the scale of the challenge. This will be increasingly relevant because future international climate policy, to facilitate decarbonization in India, must account for this subnational variability instead of treating all of India uniformly. Consequently, the differences we quantify have implications for India's decarbonization efforts as it aims to increase renewable energy by 2030, meet net-zero emissions by 2070, and ensure a just energy transition for coal-dependent states in eastern India. ■



Flow of CO_2 emissions from electricity generated and consumed in India. States on the left emit CO_2 when generating electricity for states on the right. Large states, such as Maharashtra, Uttar Pradesh, or Tamil Nadu, emit for electricity consumed within their borders. However, coal-mining states, like Chhattisgarh or Odisha, export much power and associated emissions for consumption outside their borders.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: iazevedo@stanford.edu
Cite this article as S. Sengupta *et al.*, *Science* **378**, eabh1484 (2022). DOI: 10.1126/science.abh1484

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abh1484

RESEARCH ARTICLE

AIR POLLUTION

Subnational implications from climate and air pollution policies in India's electricity sector

Shayak Sengupta[1], Peter J. Adams[1], Thomas A. Deetjen[2], Puneet Kamboj[3], Swati D'Souza[3], Rahul Tongia[1,3], Inês M. L. Azevedo[4]*

Emissions of greenhouse gases and air pollutants in India are important contributors to climate change and health damages. This study estimates current emissions from India's electricity sector and simulates the state-level implications of climate change and air pollution policies. We find that (i) a carbon tax results in little short-term emissions reductions because there is not enough dispatchable lower emission spare capacity to substitute coal; (ii) moving toward regional dispatch markets rather than state-level dispatch decisions will not lead to emissions reductions; (iii) policies that have modest emissions effects at the national level nonetheless have disparate state-level emissions impacts; and (iv) pricing or incentive mechanisms tied to production or consumption will result in markedly different costs to states.

India is the world's third-largest economy and power producer (*1*, *2*). Electricity demand may double or triple by 2030 because India still has fairly low per capita electricity consumption (*3*, *4*). Such growth, although providing incredibly important energy services and improving quality of life, will exacerbate climate change effects resulting from the emissions of greenhouse gases (GHGs) and premature mortality because of air pollutant emissions. India's total GHG emissions already rank third in the world (*5*, *6*), but per capita GHG emissions remain low. Nearly all of India is breathing polluted air, and about a million people die prematurely every year as a result of outdoor air pollution (*7*–*18*), with premature deaths projected to triple by 2050 if no action is taken to reduce air pollution (*12*).

The power sector in India contributes to both air pollution and climate change because coal-fired power stations without air pollution–control technologies contribute to the bulk (>70%) of electricity generation (*19*–*21*). Indian coal electricity generation contributes to ~40% of India's total GHG emissions (*22*). These coal plants also contribute to 50% of total sulfur dioxide (SO_2) emissions and 40% of total nitrogen oxide (NO_X) emissions. These pollutants lead to the formation of secondary $PM_{2.5}$ (particulate air pollution, specifically particulate matter <2.5 μm in diameter), resulting in premature mortality. Planned expansions of coal generation capacity will exacerbate these effects (*23*–*30*).

[1]Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA, USA. [2]Center for Electromechanics, University of Texas at Austin, Austin, TX, USA. [3]Centre for Social and Economic Progress (formerly Brookings India), New Delhi, India. [4]Department of Energy Resources Engineering, Woods Institute for the Environment, and Precourt Energy Institute, Stanford University, Stanford, CA, USA.
*Corresponding author. Email: iazevedo@stanford.edu

India has planned sulfur emissions–control regulations (*31*–*33*), the increased penetration of renewable energy (*34*), and market reforms to coordinate and economically dispatch electricity-generating units on a limited basis at the national level as opposed to the state level (*35*). Currently, each state in India schedules and dispatches its own power, largely through long-term power purchase agreements between generators and distribution companies (which govern 90% of power transactions) (*36*). Eighty-two percent of renewable capacity is concentrated in 8 of 32 states and territories. Most of this renewable capacity is found in southern and western India, where wealthier states are located. Conventional hydro capacity is located predominantly in Himalayan northern and northeastern India. Coal capacity exists throughout the country, but the cheapest coal plants are located near coal mines in a handful of poorer, eastern states (*37*, *38*). This leads to electricity generation mixes varying widely by state (*36*). States are responsible for delivering power to consumers (*36*) and own a plurality of monitored conventional capacity (*37*). Likewise, both the central and state governments have overlapping jurisdictions over the power sector. This institutional framework leads to heterogeneous consequences from centralized nationwide policies.

Previous efforts have used detailed techno-economic modeling to evaluate emissions reductions from power plant air pollution control (*39*–*41*), greater renewable energy (*42*–*50*), emissions-minimizing power sector operations (*51*), and market reforms (*35*, *52*–*56*). Some of these previous efforts have explored state-level, spatial differences in policy outcomes, with some emphasis on current Indian wholesale power market structure (*42*, *43*, *45*, *49*, *50*). However, most efforts have largely focused on a future Indian grid free of current institutional constraints, where power is dispatched in a centralized manner. The locations where electricity is generated versus where the power is used also differ, which could result in important differences for policies that focus on consumption- versus production-based emissions—an aspect that has not been quantified in previous studies. Moreover, no study has explored the potential near-term spatial or state-level emissions implications from policy interventions in Indian power sector operations if current institutional and market practices remain in the future. In this work, we develop and present a reduced-order dispatch model of Indian power generation to assess CO_2 and SO_2 emissions impacts of policy interventions to address this gap. We focus on short-term effects for the current Indian grid and analyze how policies could induce spatial differences in attributable emissions between states in India. We present state-level, production-based, and consumption-based average annual emission factors for India arising from power sector operations. Although previous work has used dispatch or capacity expansion modeling to simulate Indian power generation (*35*, *42*–*47*, *49*–*52*, *56*), this work uses a flexible, computationally simplified method. It reflects the real institutional and market organization of the Indian grid, where each state self-schedules its electricity to meet its demand from a respective portfolio of long-term contracts with generators. We first evaluate the model's simulations against reported generation data (figs. S6 to S13). We then use the model to simulate national policy intervention scenarios: carbon taxes, stricter sulfur-control regulations, and regional scheduling and dispatch among groups of states.

Estimating the emissions of GHGs and criteria air pollutants from the Indian power sector

Currently, dispatch practices in India are a use of both economic dispatch based on variable costs of generation and heuristics based on historical practices. Heuristic practices vary by state, with some regional and national coordination (*36*, *52*).

We model the Indian system as follows. First, we identify and characterize the electricity generation fleet in India. We construct a database of all nonvariable renewable generators in India using publicly available data on capacity greater than 25 MW and unit-specific heat rates (*37*, *57*, *58*) for various years between 2014 and 2018. We choose the most recently reported or modeled heat rate. We fill in missing heat rates with a log-fit of existing heat rates for units as a function of capacity differing by coal and gas units (fig. S18).

We then combine this dataset with estimates of operating and maintenance (O&M) and fuel costs of fossil fuel generation as follows:

(i) For coal generators, we calculate production-weighted variable cost of power with the Government of India's coal dispatch database (*59*), which reports grade-wise coal amounts sold to individual power stations. We combine these amounts with grade-wise fixed prices of coal from Coal India (the state-owned coal monopoly) and state-wise coal transport costs (*38*, *60*). For plants without any reported sold coal amounts, we fill in using state-wise and ownership-wise (central, state, or private) median calculated variable cost of power. Calculated variable costs of power for coal units largely match 1:1 to reported variable costs (fig. S19) of power from the MERIT India database (*61*), which reports variable cost of generation according to long-term power purchase agreements between generators and states. (ii) For natural gas plants, we use a region-based approach with domestic and imported gas prices and applicable state taxes (*62*, *63*). For nuclear and hydropower plants, we assume the reported variable cost of generation in the MERIT India database (*61*).

For nondispatchable energy sources (i.e., wind and solar), we use average monthly diurnal renewable generation profiles. We do so by first disaggregating nationwide renewable generation data for 2018 to 2019 (*19*) to obtain diurnal profiles of renewable generation and then applying these profiles to actual monthly renewable generation for each state from September 2017 to August 2018 (*64*) (fig. S1).

The total demand for power in each state is estimated by decomposing total daily demand reported at the state level from POSOCO (*65*) by state-level diurnal load profiles of demand disaggregated at the monthly level from Energy Analytics Lab (*66*). The daily demand reported by state represents the power consumed within the state at the state boundary. We assume nondispatchable energy sources are "must-take," as is the case in current markets. Thus, we estimate net demand for each hour of the year for each state by subtracting average monthly diurnal renewable generation from estimated total hourly demand for a given hour.

The next step is to identify which plants are used to meet demand. If each state were operating independently, this step would be well approximated by developing a merit order dispatch curve of the India power system—i.e., computing net demand in an hour as total demand minus nondispatchable generation (e.g., wind and solar) and then ranking dispatchable, nonrenewable generators (i.e., nuclear, hydro, coal, and gas) from low to high marginal cost and dispatching plants up to the point where supply meets the hourly demand. However, the Indian grid also has interstate plants provide power to multiple states. We thus categorize each plant and unit in the system as either intrastate plants (which can only provide power to meet one state's demand) or interstate plants (which are required to provide specific fractions of their electricity generation to multiple states).

Intrastate generating plants have 100% of their generation serving the state in which they are located. Interstate plants have portions of their generation in each hour needed to serve out-of-state demand. We identify these capacity allocations for a plant to each state from the MERIT India database and capacity allocations from the Government of India's Central Electricity Authority. Our model captures 75 to 85% of installed capacity based off reported capacity allocations (*61*, *67*) (fig. S2). In fig. S3, we show how much of each state electricity demand is provided by interstate plants versus intrastate plants. A table with intrastate and interstate capacity allocations is also shown in table S1.

We then develop a merit order dispatch curve (*68*), where plants are ranked by increasing marginal costs and where the capacity that can be provided by each plant is multiplied by the capacity allocation factor, which will be one for intrastate plants and a factor between zero and one for interstate plants. This process is repeated for each hour of the year and for each Indian state, adjusting available generators with outage information from daily generation reports. We ignore transmission constraints, ramping, and minimum capacity factor capabilities of generators, and we do not explicitly model interstate electricity transfers to meet any shortfalls in generation to meet demand. We discuss the limited effect of ignoring ramping and minimum capacity factor constraints in the supplementary materials. These factors are examined ex post facto for compliance instead of as constraints. Likewise, we discuss the sensitivity of our results to shortfalls in generation to meet demand in the supplementary materials.

The treatment of hydropower production also warrants further details. Indian power operators currently use hydro capacity during nonmonsoon months (generally January through May and November through December) as marginal generators, placing gas generators less expensive than coal generators after this hydro capacity and dispatching coal and remaining gas generators by merit order (*36*). Moreover, hydro reservoirs in India serve other purposes besides power generation (e.g., as drinking water, irrigation, and flood control). On a diurnal basis, hydro generation highly correlates with net demand, which reflects hydro's load-following nature (fig. S4). Consequently, we structure the model to dispatch power according to increasing O&M costs but considering hydro before coal and gas plants. However, to incorporate hydro generation's load-following behavior and to reflect the availability of water in reservoirs, we first constrain daily hydro capacity from reported daily hydro generation (*37*). Then, we disaggregate that generation to the hourly level according to diurnal profiles for hydro generation (*19*). Finally, we compare the capacity available to produce that amount of electricity for the hour, and we scale the available hydro capacity for the hour accordingly to represent the effective hydro capacity available to run at 100% capacity for the hour.

Finally, we estimate SO_2 and CO_2 emission factors by multiplying unit heat rate for fossil plants by fuel composition. We estimate SO_2 because sulfur control has been the focus of Indian air pollution policy discussions since 2015 (*32*, *33*). We assume domestic Indian production–weighted average coal composition (*69*) for all plants and use the mass-balance approach presented by Srinivasan *et al.* (*39*). We do not represent variation in coal quality between plants owing to data quality issues from the Government of India's coal quality data. Because of discrepancies between the coal quality delivered to plants versus what Coal India charges plants, coal quality data better represent the prices that plants pay rather than what they actually use. (*70*). We assume domestic Indian coal only with no imported coal, which disproportionately is used in a handful of coastal locations. For gas plants, we assume natural gas for CO_2 emissions (*71*) and zero SO_2 emissions.

Using the procedure mentioned in these previous paragraphs, we estimate CO_2 and SO_2 emissions under current operations (i.e., our baseline emissions) by dispatching generators from low to high marginal cost to meet net demand while considering all the Indian-specific operations for inter- and intrastate plants. We run the model for each of the 32 Indian states and union territories individually. Once we establish the baseline emissions, we also test and model seven scenarios and estimate their effects on SO_2 and CO_2 emissions as follows.

Carbon tax policy at $10, $35, $50, and $100 per ton

We test the effects of a tax of $10, $35, $50, or $100 per metric ton of CO_2 on grid operations. Currently, India has a modest carbon tax on coal of ~$3 per ton of CO_2 charged as a fee per ton of coal that is implicitly incorporated into the variable cost of power from plants (*72*). The taxes we test are in addition to variable costs of power. We translate the tax to an additional cost to the plant operator using the carbon intensity of each plant. For the carbon tax scenarios, we assume $1 USD = ₹71 INR, which adds ~300% to the average variable cost of coal generation with a $100/ton tax.

Air pollution–control technology adoption

The air pollution policy represents the minimal control and costs needed to meet unimplemented

SO_2 emissions regulations in India (table S2) (*39*). As of December 2019, only 13.75 GW of capacity (~5% of monitored capacity) in India have any operational sulfur control (*31*), and we cannot verify whether installed control runs regularly because of a lack of publicly available continuous emissions monitoring data. Moreover, in March 2021, the Government of India extended implementation deadlines to 2025 (*33*), with current policy discussions highlighting whether certain regions or plants need priority because of high capital costs of sulfur-control equipment. Therefore, for simplicity, we assume no sulfur control at any plants in the base case scenario. Assumptions on control technologies and their variable costs come from Srinivasan *et al.* (*39*). For this scenario, we first calculate the percentage reduction required for each unit to meet Indian sulfur-control regulations. Then we assume the minimal, least-cost control technology required to meet that reduction from those presented in Srinivasan *et al.* (*39*). For 368 units totaling 72 GW of capacity, we assume dry limestone rejection to reduce sulfur emissions ~60% at annualized cost of ₹6000 per ton of SO_2 removed ($85 per ton). For the remaining 279 units totaling 129 GW of capacity requiring reductions >60%, we assume wet flue gas desulfurization at annual cost of ₹7000 per ton of SO_2 ($99 per ton). These costs only include reagent costs and increase capacity-weighted variable cost of power by 1 to 2%. All other associated control costs fall in fixed costs according to Indian regulations; they do not influence dispatch by variable cost of generation (*39*). We also run a scenario where we impose a $35 per ton carbon tax along with sulfur control.

Moving Indian power grid operations toward regional markets

We also test Indian states participating in the same electricity markets by region, instead of state by state. Current policy discussions in India aim to shift power dispatch to more coordination between states to share capacity and lower costs (*35*, *53*, *55*, *73*). Moreover, work modeling future Indian grids assumes more renewable generation incumbent upon regional or national dispatch to move electricity around the country (*45*, *47*–*49*). We consider each Indian region (see fig. S5 for a map of the regions)—i.e., north, south, east, west, and northeast—as separate single markets where demand needs to be met. We still maintain the same modeling assumptions regarding renewables as described before. We ignore transmission constraints, so this assessment provides an upper bound of the economic benefits that would rise from merging state markets into regional ones. We rerun the dispatch model under these new regional boundaries. For each scenario, we operate the model at hourly resolution for September 2017 to August 2018.

What are the CO_2 and SO_2 emissions from electricity generation for each state?

We find that overall emissions of CO_2 from the power sector are 820 megatons of CO_2, or an average emission intensity of 711 kg of CO_2 per megawatt-hour (MWh). SO_2 emissions amount to 6100 kilotons of SO_2, with an average emission intensity of 5.3 kg of SO_2/MWh (figs. S6 to S8). Estimated emissions and generation match the reported data for most fuels and regions within 30% (figs. S3 and S6 to S13). However, considerable heterogeneity exists between states (figs. S9 to S13 and S27 to S31). Our estimated annual wholesale O&M costs for electricity are $58.5 billion (₹4.2 lakh crore) (fig. S20), which is ~14% less than reported annual wholesale costs that state distribution utilities incurred to procure power in 2017 to 2018 (*74*).

We compute both production- and consumption-based emissions and their average emissions intensity for each state (Fig. 1, Fig. 2, and tables S3 and S4). Production-based emissions are the emissions associated with generation within

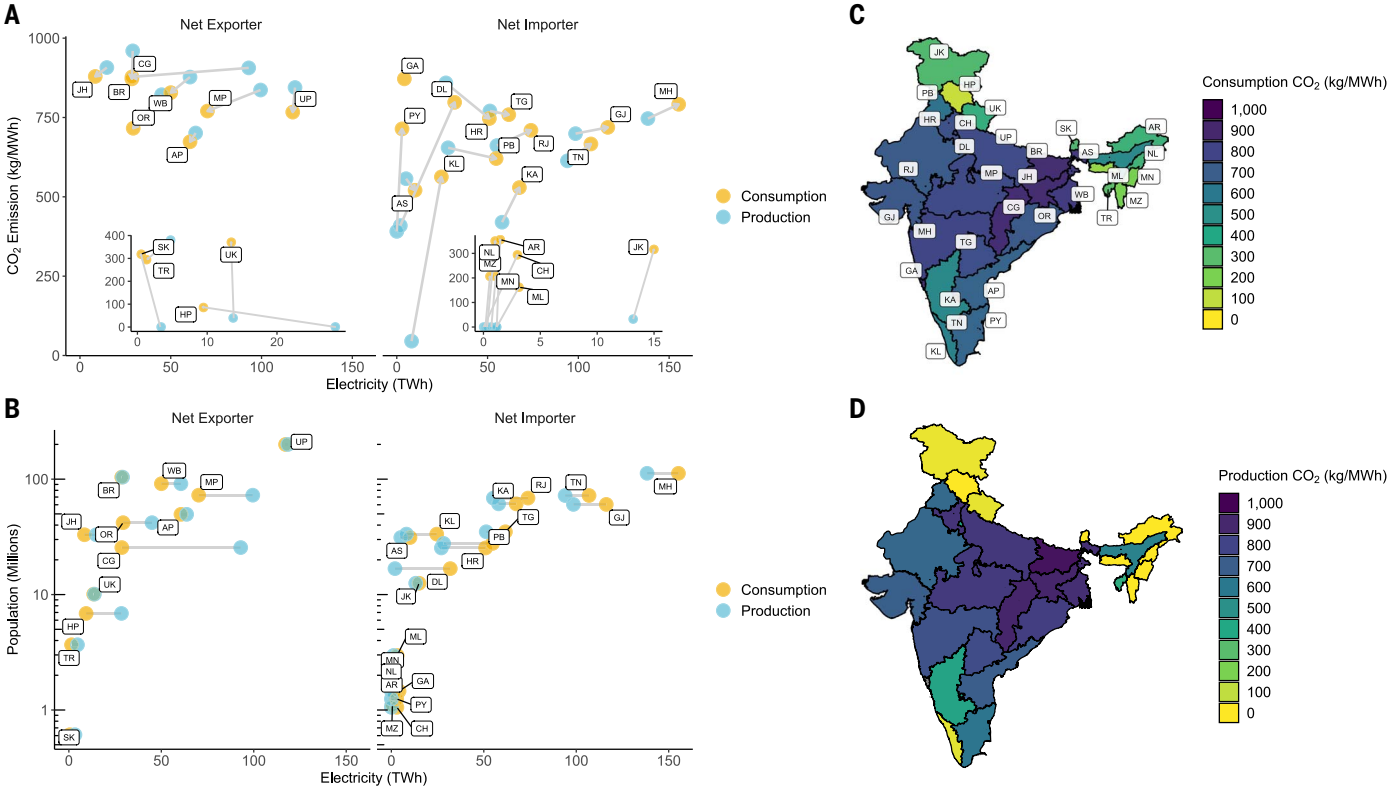


Fig. 1. Differences in electricity CO_2 emissions between states in India. (**A**) Production (blue) and consumption (yellow) carbon intensities of each state versus production (blue) and consumption (yellow) (*y* axis) versus electricity generation or consumption (*x* axis), split by states that are either net exporters or net importers of electricity. Inset shows low-demand states. TWh, terawatt-hours. State abbreviations are provided in table S3. (**B**) Population of each state versus the annual electricity production (blue) or consumption (yellow) in a state. (**C**) Map of consumption-based CO_2 emissions intensity for each state. (**D**) Map of production-based CO_2 emissions intensity for each state.

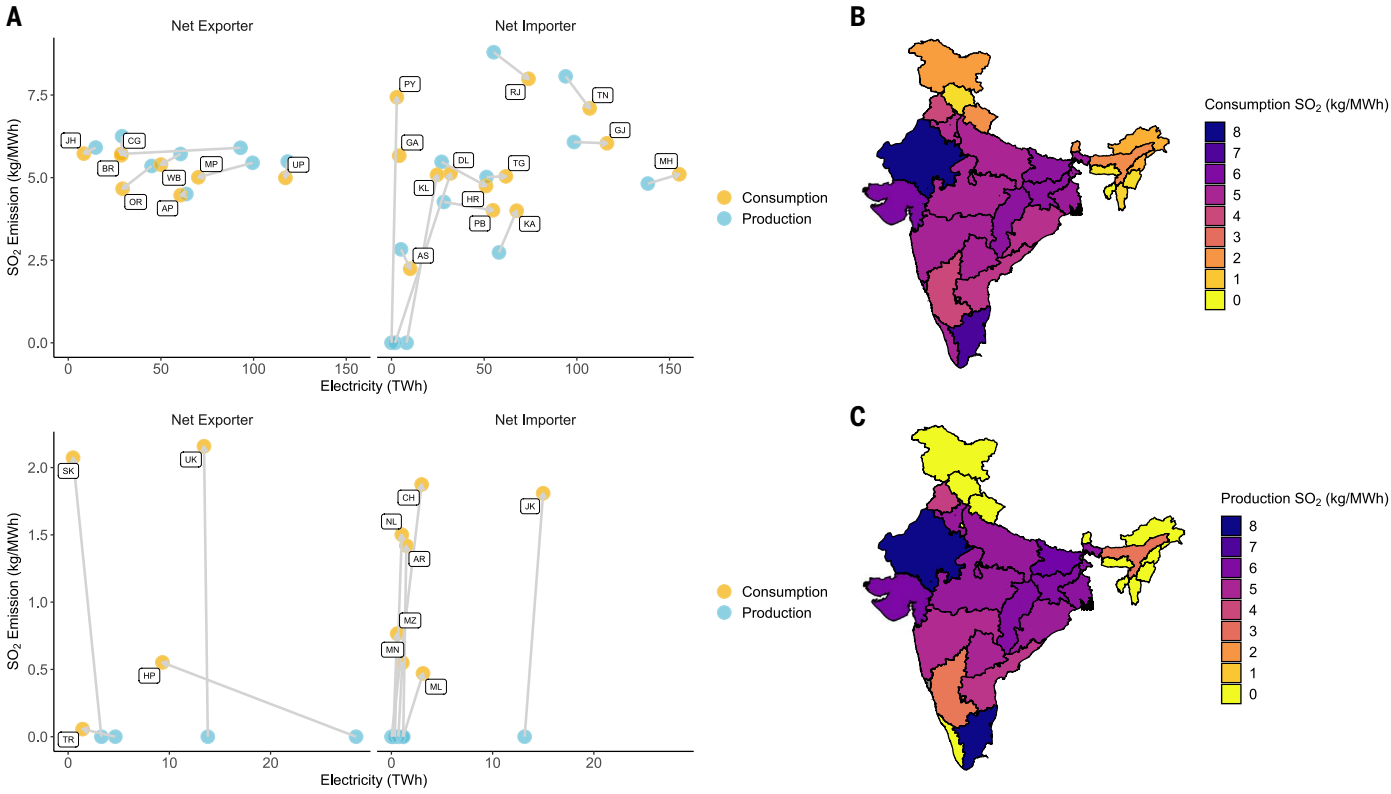


Fig. 2. Differences in electricity SO_2 emissions between states in India. (**A**) Production (blue) and consumption (yellow) sulfur intensities of each state (*y* axis) versus production (blue) and consumption (yellow) generation (*x* axis), split by states that are either net exporters or net importers of electricity. (Top) High-demand states. (Bottom) Low-demand states. (**B**) Map of consumption-based SO_2 emissions intensity for each state. (**C**) Map of production-based SO_2 emissions intensity for each state.

a state's borders only. Consumption-based emissions are based on carbon intensity of the electricity accounting for imports and exports. We separate states by net exporters (where electricity generation is higher than electricity consumption) and net importers (where electricity generation is lower than electricity consumption). See table S3 for a list of state abbreviations and full names.

Annual production-based CO_2 emissions range from nearly zero [in Arunachal Pradesh (ARP), Chandigarh (CH), Goa (GA), Himachal Pradesh (HP), Manipur (MN), Meghalaya (ML), Mizoram (MZ), Nagaland (NL), and Sikkim (SK)] to 103 megatons of CO_2 [in Maharashtra (MH)], and emissions intensity ranges from 0 to 960 kg of CO_2/MWh [with the high end of this range occurring in Bihar (BR)] (Fig. 1, A and D). Several states report zero production-based emissions because of zero-carbon capacity or because they import all of their electricity (*37*, *61*)

Production-based SO_2 emissions range from 0 kilotons [in Arunachal Pradesh, Chandigarh, Delhi (DL), Goa, Himachal Pradesh, Jammu and Kashmir (JK), Kerala (KL), Manipur, Meghalaya, Mizoram, Nagaland, Puducherry (PY), Sikkim, Tripura (TR), and Uttarakhand (UK)] to 760 kilotons [in Tamil Nadu (TN)], corresponding to emissions intensities ranging from 0 to 8.8 kg/MWh (Fig. 2A). We assume that natural gas–based capacity emits no SO_2. States with the highest SO_2 emissions have the largest coal capacity, including those with lignite-burning stations [Tamil Nadu, Gujarat (GJ), and Rajasthan (RJ)], which have a higher sulfur content than average Indian coal.

We calculate consumption-based emissions by summing the emissions from interstate and intrastate generation associated with meeting demand for a state. The highest consumption-based emissions occur in Maharashtra (120 megatons of CO_2), and the lowest occur in Mizoram (0.10 megatons of CO_2). The carbon intensity for consumption-based emissions ranges from 87 kg of CO_2/MWh (in Himachal Pradesh) to 879 kg of CO_2/MWh [in Jharkhand (JH)] (Fig. 1, A and C).

There are two distinct clusters in the spread of consumption emission factors. At the lower range of state-level emission factors are 11 low–annual demand states (~4% of total nationwide demand), primarily Himalayan and northeastern states, where considerable hydro capacity exists (Fig. 1A, inset, and Fig. 2A, bottom panel). At the minimum within this group, Himachal Pradesh reports an average emission factor of 87 kg of CO_2/MWh with Uttarakhand at the greatest at 372 kg of CO_2/MWh. For SO_2, Tripura reports the lowest at 0.1 kg of SO_2/MWh, and Uttarakhand reports the highest at 2.2 kg of SO_2/MWh. Within this cluster of low-demand states, those with higher CO_2 emission factors tend to depend on a combination of coal and gas paired with hydro. Those with higher SO_2 emission factors depend on coal more than gas. The second cluster of states are the remaining 21 states simulated, forming 96% of total annual demand. Among this group, average consumption annual emission factors vary considerably, with Assam (AS) and Karnataka (KA) at the lower end of this group (521 kg of CO_2/MWh and 2.2 kg of SO_2/MWh; 530 kg of CO_2/MWh and 4 kg of SO_2/MWh) and with the highest emission factor among all states coming from Jharkhand (879 kg of CO_2/MWh). The highest SO_2 emission factor is in Rajasthan (8 kg/MWh), which depends in part on lignite plants. Although Karnataka, Madhya Pradesh (MP), Tamil Nadu, Maharashtra, Gujarat, Rajasthan, Andhra Pradesh (AP), and Telangana (TG) have the highest amounts of renewable capacity, they also use considerable amounts of coal capacity, putting them in the highest cluster of states.

What are the effects of a carbon tax?

Relying on a carbon tax to reduce CO_2 emissions in the Indian power system will prove

ill-founded in the short term. Even under a \$100 per ton of CO_2 tax (see Fig. 3A and fig. S16 for remaining tax levels), under the current electricity generation fleet, only 12 states would see reductions in their production emissions intensity larger than 5%: Assam (6% decrease), Gujarat (9%), Kerala (9%), Punjab (10%), Arunachal Pradesh (11%), Meghalaya (11%), Jammu and Kashmir (12%), Manipur (15%), Chandigarh (16%), Nagaland (18%), Mizoram (23%), and Uttarakhand (27%). These states account for ~20% of annual nationwide demand and 15% of electricity production. We find that the annual average nationwide CO_2 (production) emission factor decreases from 711 to 686 kg/MWh, a decrease of 4%. We show the effect of carbon taxes on annual plant capacity factor (how often a plant runs) in the supplementary materials (figs. S23 to S26). Total annual wholesale marginal costs increase 15 to 150% with progressively increasing carbon taxes, and increases in average marginal costs disproportionately affect eastern coal-mining states in all but the highest carbon taxes of \$100 per ton (fig. S21). There may still be very valid reasons to implement a carbon price even if the short-term benefits are negligible. Doing so may signal to the market the needs for low carbon generation in the long term.

There is an important co-benefit from a carbon tax because it induces reductions of SO_2. Under a \$100 per ton of CO_2 carbon tax (Fig. 3B), total SO_2 emissions decrease 11% from 6100 to 5400 kilotons in this scenario, with 47 plants increasing their SO_2 emissions, 20 seeing no change, and 95 incurring a decrease in SO_2 emissions. Several plants that see decreases in SO_2 emissions are lignite-burning plants in Tamil Nadu, Rajasthan, and Gujarat, with higher heat rates (lower efficiency) and higher sulfur emissions. Although these plants are typically next to lignite mines, there is a paucity of lignite transport costs. We assume similar relationships for transport costs—i.e., plants further away from eastern coal-mining regions have higher transport costs. Consequently, the carbon tax penalizes these lignite plants for both lower efficiency and location away from coal mines. The benefits to reducing premature mortality from air pollution associated with SO_2 emissions will depend on how these emissions would relate to changes in $PM_{2.5}$ concentration and the resulting changing in exposure. In any case, if we assume a very simplified order-of-magnitude effect and assume that this decrease in SO_2 results in an average nationwide decrease in $PM_{2.5}$ concentration of ~1 $\mu g/m^3$ and an average population density, this would translate to ~10,000 avoided premature deaths per year (*27*).

What are the effects of the proposed air pollution–control policies?

For sulfur control, we assume the implementation of two control technology options—dry limestone injection or wet flue gas desulfurization—from Srinivasan *et al.* (*39*), which analyzed an entire suite of possible control technologies. Overall, we see a 79% decrease in annual nationwide sulfur emissions from 6100 kilotons in the base case to 1300 kilotons in the sulfur-control scenario (Fig. 4A). Annual costs increase by ~1% (fig. S21). These decreases represent the minimal control needed to meet upcoming Indian SO_2 emissions standards for coal generators (table S2). Likewise, minimal nationwide control yields little change (<5%) in plant load factor (PLF)—i.e., the capacity factor of plants (Fig. 4B). Only 19 of 162 coal plants dispatched see appreciable changes in PLF, with 88% of plants showing no change. These 19 plants are located throughout the country.

Our base case sulfur emissions estimate and results from sulfur controls are consistent with previous analyses of sulfur control at Indian power stations. Srinivasan *et al.* (*39*) have reported 95% reductions to 650 kilotons, with SO_2 emission factors decreasing from 7.9 to 0.4 kg/MWh with more stringent control under a wider range of control technologies. Likewise, previous analyses estimating total SO_2 emissions from the Indian power sector have ranged from 3500 to 10,100 kilotons (*24*, *57*, *75*–*77*).

Should India consider regional markets as a means to reduce emissions?

Figure 5 shows the changes between the regional dispatch scenario emissions and current emissions. Figure 6 shows the absolute production CO_2 and SO_2 emission factors currently (business-as-usual scenario) and in the regional dispatch scenario. We show the corresponding consumption emission factors in fig. S17. Regional dispatch means that each region would dispatch its plants to meet demand in the region, rather than having each state doing so separately. A regional dispatch approach would increase, rather than decrease, the average nationwide CO_2 emission factor from 711 to 720 kg/MWh. Total costs decrease by ~6%, consistent with previous estimates (*35*, *55*). However, the state-level effects that this regional pooling produces are disparate in terms of emission factors (Fig. 5A) and costs (fig. S22). In Fig. 6A, 14 of 32 states see a decrease in average production emission factors (ranging from 1 to 38% reduction depending on the state), and 9 states see an increase in carbon intensity, with the highest increases for states with the lowest emission factors in the base

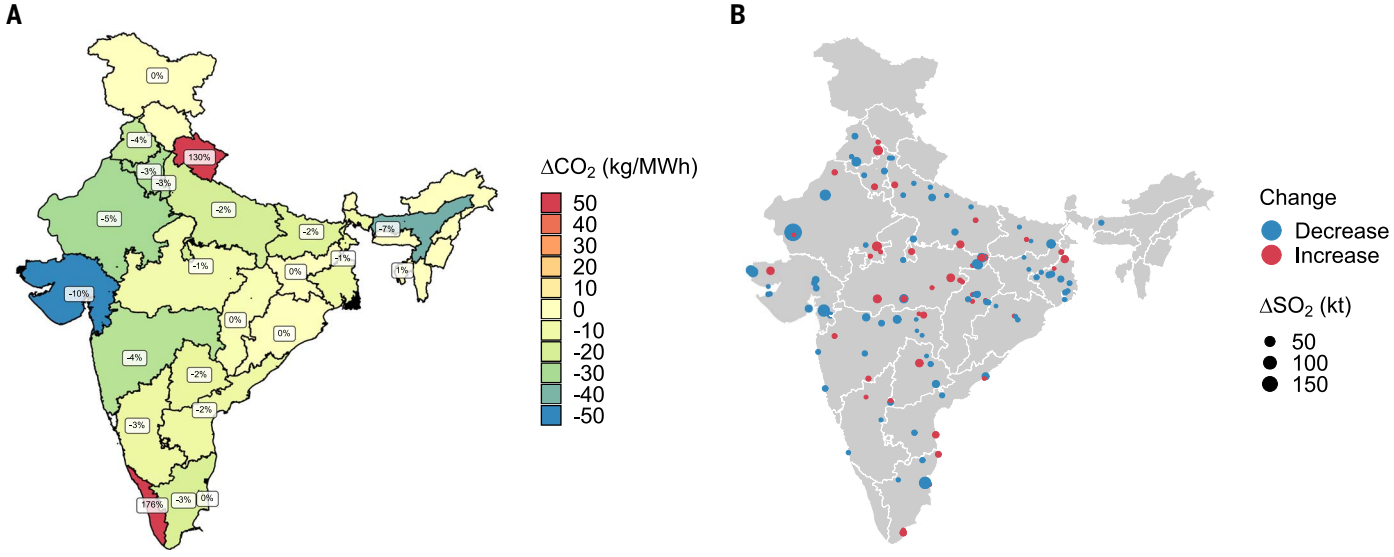


Fig. 3. Impacts of high carbon taxes on the Indian power sector. (**A**) Changes in average annual production CO_2 emission factors by state for a \$100 per ton of CO_2 tax. (**B**) Changes in SO_2 emissions induced from a \$100 per ton of CO_2 tax. kt, kiloton.

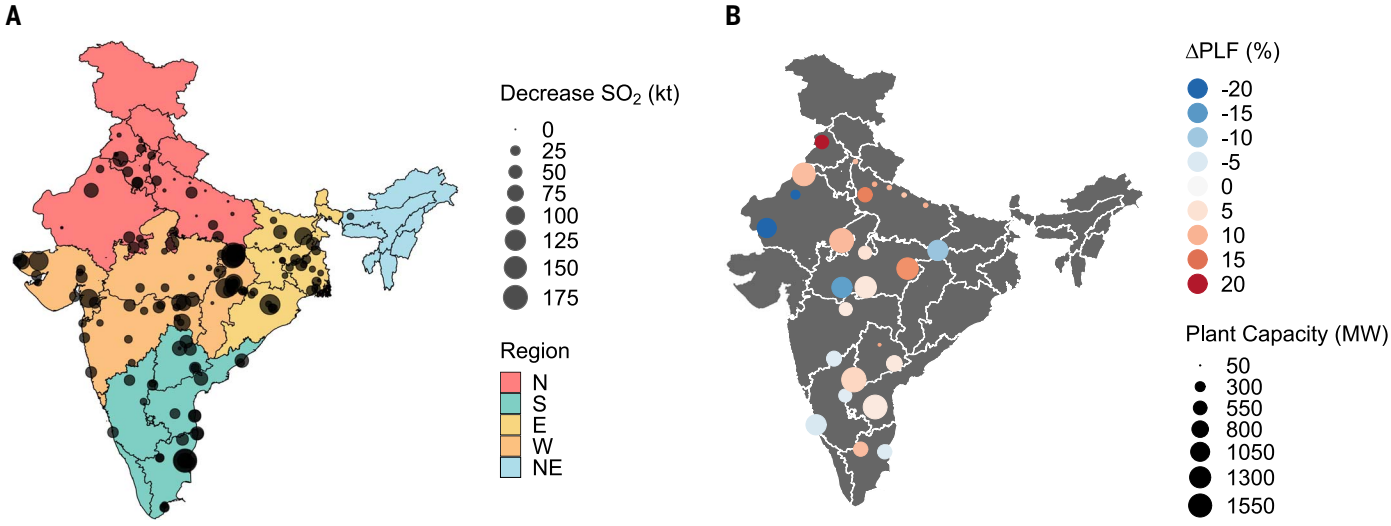


Fig. 4. Impacts of sulfur-control policies on the Indian power sector. (**A**) Decrease in SO_2 emissions from implementing minimal sulfur control to meet Indian emissions norms when compared with baseline emissions. (**B**) Changes in coal PLFs under the SO_2 emissions policy versus the baseline.

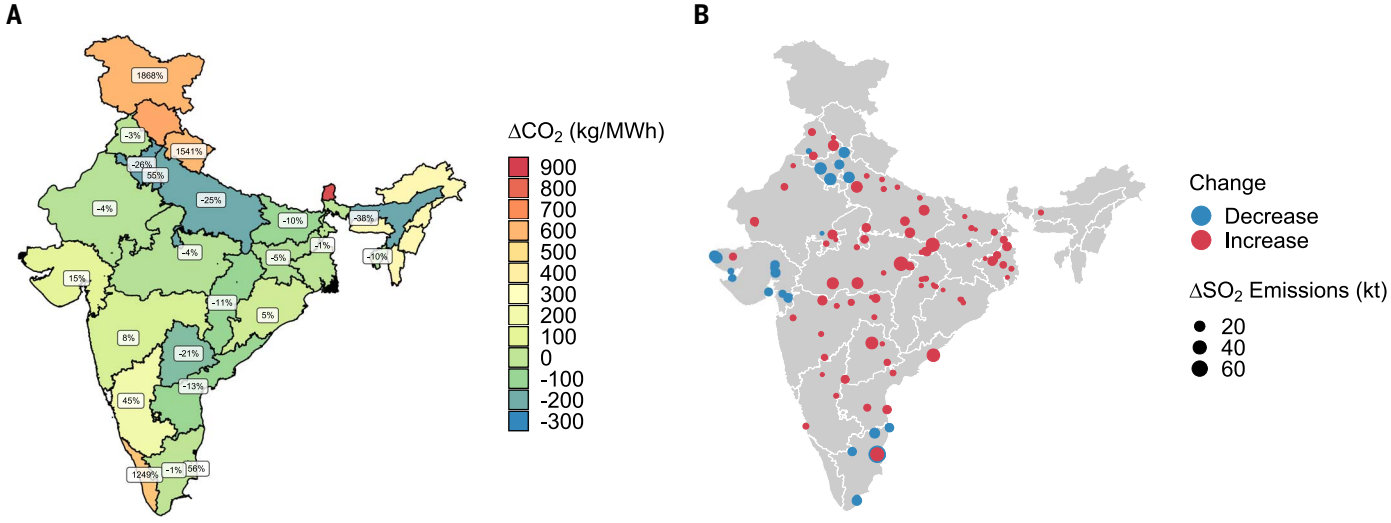


Fig. 5. Impacts of regional coordination between states to dispatch power in India. (**A**) Changes in average annual production emission factors by state for regional dispatch. Labels show percentage changes, where applicable. (**B**) Changes in SO_2 emissions induced from regional dispatch.

case (e.g., Jammu and Kashmir, Himachal Pradesh, and Sikkim). The SO_2 effects of this simulation show, once again, SO_2 emissions concentrated in different spatial patterns (Fig. 5B). Nationwide SO_2 emissions increase to 6200 kilotons in this scenario compared with the base case scenario, an increase of 2%. The number of plants seeing increases or decreases compared with the base case differs. Of 162 plants, 97 see increases in SO_2 emissions, 31 see decreases, and 34 see no change. However, clear clusters emerge. Plants that are the furthest away from coal-mining areas (e.g., Haryana, Gujarat, and Tamil Nadu) see decreases in SO_2 emissions, whereas plants closer to mines see increases as a result of closer plants having lower variable costs with total fuel costs heavily dependent on coal transportation costs.

Key findings and policy considerations

Indian power sector policies, operations, capacity mix, and generation vary by Indian state. The increased penetration of renewable energy, the need to control air pollution, and planned market reforms warrant analysis beyond national aggregate metrics of CO_2 and SO_2 emissions from Indian power generation. Our nationwide average emission factors agree with the previous national estimates (*45*, *49*, *50*, *78*, *79*).

Our results suggest that designing climate mitigation strategies and emissions reductions programs and goals on the basis of states' geographical boundaries may miss these complex interactions between where electricity is produced and where it is used. For example, the effect of demand-side emissions reduction policies (such as energy efficiency) or the impact of future loads (such as increased demand for air conditioning) that use production-based emission factors derived from generators located within geographic boundaries of states will yield inaccurate results. Although for some states with the largest demands, such as Maharashtra, Tamil Nadu, or Gujarat, the difference between consumption-based and production-based emission factors is less than ±10%, for other states with large demands, such as Karnataka or Uttar Pradesh, the difference can be in excess of ±10 to 20%. The difference is most pronounced in states with smaller demand, which are more likely to import electricity from neighboring states. This increases their consumption-based emission factors over their production-based emission factors (i.e., Himalayan or northeastern states).

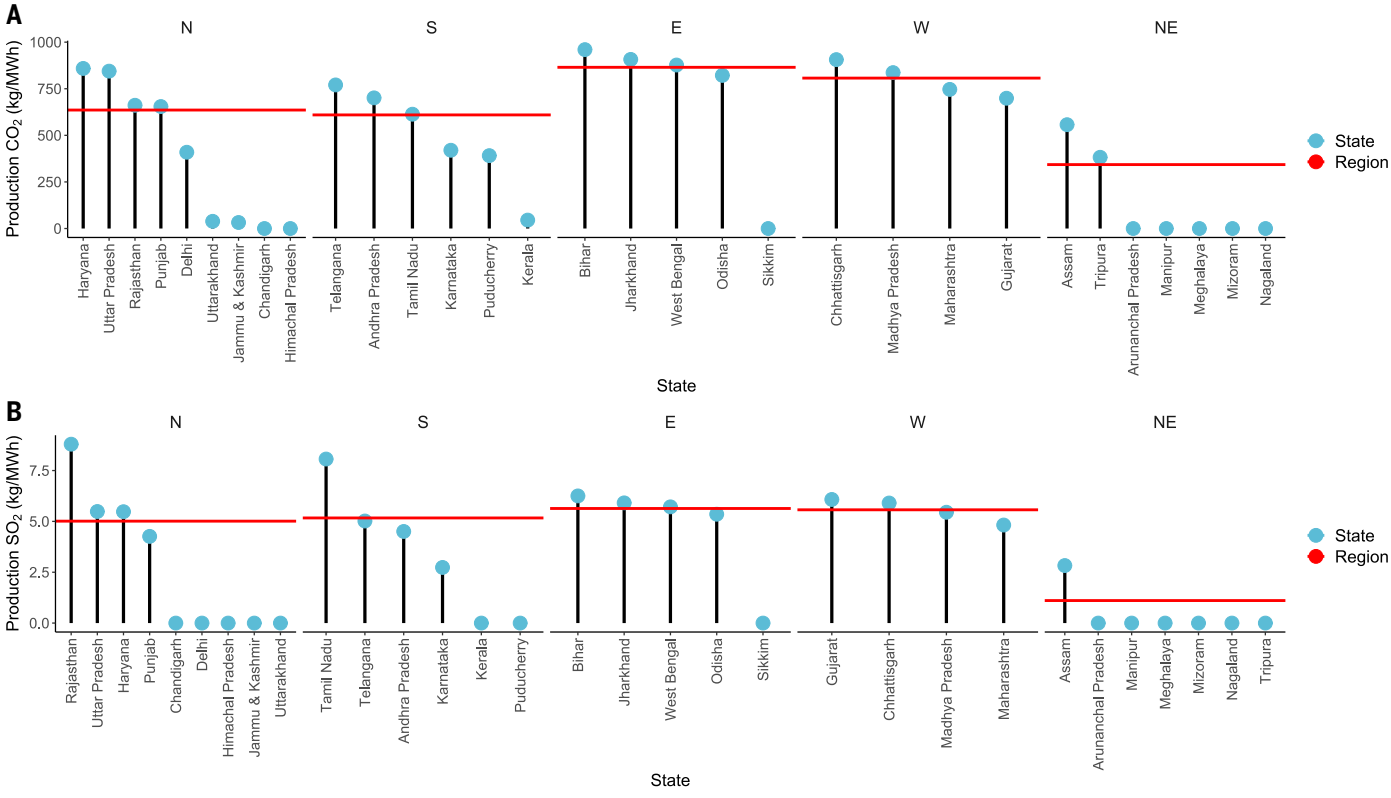


Fig. 6. State-by-state impacts of regional coordination between states to dispatch power in India. (**A**) Average annual CO_2 production emission factors by state for region and baseline (state) dispatch scenarios. (**B**) Average annual SO_2 production emission factors by state for region and state dispatch scenarios.

State-level, consumption-based emission factors provide more detailed and accurate information for policy analyses compared with both national-level emission factors and regional-level factors because, to meet demand, individual states enter contracts with generators from multiple owners located inside and outside a state.

Pricing or incentive mechanisms based on production or consumption will result in markedly different costs to a state. When mechanisms are tied to production, net exporter states that generate higher-emissions electricity (i.e., coal-mining eastern areas) will face higher costs compared with net exporter states that generate lower-emissions electricity (i.e., hydro-rich Himalayan areas). Likewise, mechanisms tied to consumption will likely increase costs to net importer states that consume higher-emissions electricity (i.e., large-demand states with high populations and large cities) compared with net importer states that consume lower-emissions electricity (i.e., small-demand states with low populations).

Moreover, on the supply side, production-based emission factors with states as net power importers and exporters means that efforts to increase shares of zero-emission generation in general would provide wider reach if put into high-emission, net exporter states (e.g., Odisha, West Bengal, Jharkhand, and Chhattisgarh). Transmission networks are situated to deliver this power from these states to net importer states. When coal plants in these states retire, their accessibility to transmission make them well situated for renewable generation as a replacement. Pai *et al.* (*80*) have found that nearly all areas dependent on coal mining in India have high solar energy potential.

Electricity decarbonization and emissions reduction efforts in India must show an appreciation for the scale of the challenge. Electricity production and consumption and associated emissions between Indian states show orders-of-magnitude differences tied to population, akin to the differences observed between different countries around the world. This will be increasingly relevant because future international climate policy, to facilitate decarbonization in India, must account for this subnational variability instead of treating all of India uniformly.

In addition to the business-as-usual scenario, this analysis quantifies and defines the spatial CO_2 and SO_2 effects that would result from different policy interventions in the current Indian power sector. These policy interventions include (i) a carbon tax in addition to the current implicit carbon tax charged to mined coal; (ii) minimum sulfur control to meet current, but unimplemented, Indian regulations; and (iii) regionally coordinated dispatch instead of dispatching plants at the state level, consistent with the Government of India's plans to coordinate interstate dispatch (*35*, *53*, *56*).

We recognize limitations with our estimates, which will reflect short-term conditions that derive from the current electricity-generation capacity mix in India. By focusing on the short term, as opposed to the longer term, our results highlight current spatial differences that are the result of current institutional and market constraints. Regardless of specific policy instruments, such as a carbon tax, over time there will be a rise of cleaner technologies, and, in fact, there are indications that greenfield coal power plant deployments will be few if not nil. This is because of sufficient overcapacity in the very short term as well as under-construction plants that should suffice through the end of the decade. This highlights the importance of existing coal capacity by 2030 (*50*, *81*).

Of the policy interventions we evaluate, the carbon tax is most sensitive to our assumption of a fixed set of available generation capacity. A carbon tax, even at a fairly high level of $100 per ton of CO_2, fails to yield substantial CO_2 emissions reductions in the current Indian grid in the short term. This is in addition to the implicit carbon tax of ~$6 per ton of domestic and imported coal that India currently charges for coal, which translates to a little

more than $3 per ton of CO_2 (*72*). We find that the tax required on a coal generator to achieve parity in variable cost with a gas generator would be ~$66 per ton of CO_2, but there is simply not enough natural gas that could be used to displace large volumes of coal. In the long term, a carbon tax would likely spur future investment in lower carbon generation capacity in India, lowering both electricity GHG and sulfur emissions over time. However, unless policies and programs geared toward renewables coupled with storage and nuclear or natural gas occur in parallel, the effect of the tax in the first few years would be just to increase electricity costs to producers and thus electricity bills to consumers. These increased costs would disproportionately affect poorer, coal-heavy eastern states. Public planning documents that inform Government of India policy project ~50% generation from coal by 2030 despite renewable capacity growth (*81*, *82*), and detailed dispatch modeling using this planned future capacity finds similar levels of coal generation in the future (*45*). When modeling least-cost capacity planning to 2030 and 2047, in an absence of an additional carbon tax, previous analyses have found coal generation standing at ~48 and ~41%, respectively (*49*, *50*).

The implementation of sulfur controls will likely result in large reductions of SO_2 emissions, with the important outcome of reducing the current premature mortality associated with air pollution in India. Our results show that most plants see minimal changes in power load factors (capacity factor) and how often a plant runs from the implementation of sulfur control. Sulfur control imposes minimal variable-cost increases, which minimally influence the order in which plants dispatch to produce electricity. Moreover, we assume that all plants implement sulfur control. If policymakers target specific plants in a regional manner, as currently proposed (*33*), the potential for shifting emissions away from target areas to other areas is possible because of the statewise power dispatch in India and the additional marginal cost of sulfur control, penalizing plants in the dispatch order. Expected future life span and output would further skew the capital costs, which could ultimately lead to nonuniform deployment, with the extreme example of some plants simply shutting down. Policy measures to date do not factor in geographic effects of pollution, but the Government of India has proposed such a graded plan for emissions-control equipment based on location. This approach acknowledges the spatial differences in pollution burden from each plant but fails to account for the secondary formation of air pollution from precursor gases emitted by plants (*33*). We should note that our variable costs of control (₹20 to ₹100 per megawatt-hour) are less than the ₹50 to ₹200 per megawatt-hour penalties proposed by the Government of India for noncompliant plants, which suggests that plants are better off adopting control technology to avoid penalties in the dispatch order. PLF changes can guide capital investment decisions because capital costs contribute to most sulfur-control costs. Specifically, for plants that see decreases in power load or capacity factors, capital costs may exceed the power load factors needed to recoup these costs during plant operation. It remains to be seen whether Indian regulators allow a pass-through of such costs even in lowered–power load factors cases. This may affect which plants ultimately do or do not undergo sulfur-control installation (*31*). No clear patterns about plant age, size, or location emerge for the plants that do see appreciable changes in PLF in our sulfur-control scenario.

As has occurred in many regions around the world, India could consider the creation of regional electricity markets to improve the efficiency of the system and decrease costs. However, that would not necessarily align with a decrease in emissions. Our simulations suggest that dispatching plants at the regional level rather than at the state level would lead to a small increase in both SO_2 and CO_2 emissions. Regionally coordinated dispatch would also impose changes in the spatial patterns of SO_2 and CO_2 emissions by shifting emissions from distant plants even more to plants closer to eastern coal-mining regions in Chhattisgarh, Odisha, Jharkhand, Bihar, and West Bengal. Because of intensive coal mining and power generation, these areas already face a disproportionate burden of pollution from coal (*23*). Likewise, SO_2 emissions become more spatially dispersed, with the number of plants seeing SO_2 emissions increases greater than those seeing decreases. This behavior is consistent with findings from Kamboj and Tongia (*38*), who have found coal transport costs to predominantly determine the variable cost of electricity for Indian power stations. A regional market would penalize plants in states with the highest transport costs—e.g., Gujarat, Haryana, and Tamil Nadu.

Notably, our model ignores any explicit transmission constraints that would limit interstate electricity trade in regional dispatch. Our estimates likely give an upper bound on the shifting of emissions and generation to coal-mining eastern states because transmission constraints would limit how much generation from these areas would substitute generation in further-away areas. However, previous modeling suggests that the spatial patterns in these results would qualitatively apply to a future Indian grid. Spencer *et al.* (*45*), who modeled a nationwide integrated 2030 Indian grid with expanded renewables (with ~50% coal generation) and transmission constraints, have found that coal-heavy states consistently remain net power exporters during representative times of day, especially at night when there is no solar output. Likewise, Rose *et al.* (*49*) modeled least-cost electricity capacity expansion with transmission constraints in India to 2017 and 2047. They have found that elevated emissions in 2017 and nearly all emissions in 2047 concentrated to eastern coal-mining states because coal power remains cost competitive in these areas. Lastly, Abhyankar *et al.* (*50*) modeled least-cost pathways to meet 2030 renewable energy targets in India with current state-wise dispatch. They have found that eastern coal-mining areas export most coal generation. Consequently, a regionally or nationally coordinated Indian grid will allow greater shares of renewables to move about the country to lower total emissions. However, coal-mining regions with cheaper coal plants will remain generation and emissions hotspots if the interstate disparity in the location of renewable and coal capacities remains. One way to make coal power less cost competitive in the future in eastern areas would be through a carbon tax, as we explored here, but policy-makers would need to carefully design such a policy to ensure that poorer consumers in these states would not see increased electricity costs. The regressive nature of any such tax could be overcome through redistribution mechanisms, but historical poor achievements highlight the challenge for a developing country with a high fiscal deficit. The ₹400 per ton "coal cess" began life as a "clean energy and environment cess" but, since 2017, has been used purely for budgetary support by the Government of India (*83*).

In addition to the limitations discussed with each scenario, we identify several other limitations in our modeling and analysis. First, our model does not capture 100% of generating capacity, instead reflecting 75 to 85% of capacity tied to long-term power purchase agreements (fig. S2), which govern 90% of power transactions in India (*73*). Our results consequently predict 11% lower total CO_2 emissions in our base case than the Government of India estimates (*79*) and an uncertainty range of 693 to 721 kg/MWh in the national CO_2 emission factor that we derive. We explain in detail the uncertainty in our results associated with this assumption in the supplementary materials. A second limitation we recognize is that we only consider CO_2 and SO_2 emissions, despite the Indian power sector being a large source of NO_X and primary (directly emitted) $PM_{2.5}$ emissions. Although emissions of NO_X are dependent on combustion conditions, we use a mass balance approach to derive plant-specific emission factors (*39*) because of the lack of continuous emissions monitoring data for India. Adopting the same approach to derive NO_X and $PM_{2.5}$ emissions would qualitatively yield results similar to those for SO_2 in this analysis.

Our analysis shows that policies that have modest or negligible emissions impacts at the aggregate, national level nonetheless have disparate, state-level, spatial emissions and cost effects. Consequently, the differences that we quantify have implications for India's decarbonization efforts as it aims to increase renewable energy by 2030 and meet net-zero emissions by 2070 while ensuring a just energy transition for coal-dependent states in eastern India (*80*, *84*).

REFERENCES AND NOTES

1. World Bank, GDP, PPP (current international $) (2018); https://data.worldbank.org/indicator/ny.gdp.mktp.pp.cd.
2. BP, "Statistical Review of World Energy" (2020); https://www.bp.com/content/dam/bp/business-sites/en/global/corporate/pdfs/energy-economics/statistical-review/bp-stats-review-2020-full-report.pdf.
3. S. Ali, "The future of Indian electricity demand: How much, by whom and under what conditions?" (Brookings India, 2018); https://www.brookings.edu/research/the-future-of-indian-electricity-demand-how-much-by-whom-and-under-what-conditions/.
4. T. Spencer, A. Awasthy, "Analysing and Projecting Indian Electricity Demand to 2030" (TERI, 2019); https://www.teriin.org/sites/default/files/2019-02/Analysing%20and%20Projecting%20Indian%20Electricity%20Demand%20to%202030.pdf.
5. World Resources Institute, CAIT Climate Data Explorer (2019); https://www.wri.org/data/cait-climate-data-explorer.
6. J. Timperley, "The Carbon Brief Profile: India" (Carbon Brief Ltd., 2019); https://www.carbonbrief.org/the-carbon-brief-profile-india.
7. S. D. Ghude *et al*., Premature mortality in India due to $PM_{2.5}$ and ozone exposure. *Geophys. Res. Lett.* **43**, 4650–4658 (2016). doi: 10.1002/2016GL068949
8. L. Conibear, E. W. Butt, C. Knote, S. R. Arnold, D. V. Spracklen, Stringent Emission Control Policies Can Provide Large Improvements in Air Quality and Public Health in India. *Geohealth* **2**, 196–211 (2018). doi: 10.1029/2018GH000139; pmid: 32395679
9. J. S. Apte, J. D. Marshall, A. J. Cohen, M. Brauer, Addressing Global Mortality from Ambient $PM_{2.5}$. *Environ. Sci. Technol.* **49**, 8057–8066 (2015). doi: 10.1021/acs.est.5b01236; pmid: 26077815
10. S. Chowdhury, S. Dey, Cause-specific premature death from ambient $PM_{2.5}$ exposure in India: Estimate adjusted for baseline mortality. *Environ. Int.* **91**, 283–290 (2016). doi: 10.1016/j.envint.2016.03.004; pmid: 27063285
11. H. Guo *et al*., Source contributions and potential reductions to health effects of particulate matter in India. *Atmos. Chem. Phys.* **18**, 15219–15229 (2018). doi: 10.5194/acp-18-15219-2018
12. GBD MAPS Working Group, "Burden of Disease Attributable to Major Air Pollution Sources in India. Special Report 21" (Health Effects Institute, 2018); https://www.healtheffects.org/publication/gbd-air-pollution-india.
13. E. E. McDuffie *et al*., Source sector and fuel contributions to ambient $PM_{2.5}$ and attributable mortality across multiple spatial scales. *Nat. Commun.* **12**, 3594 (2021). doi: 10.1038/s41467-021-23853-y; pmid: 34127654
14. J. Lelieveld, J. S. Evans, M. Fnais, D. Giannadaki, A. Pozzer, The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature* **525**, 367–371 (2015). doi: 10.1038/nature15371; pmid: 26381985
15. A. J. Cohen *et al*., Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: An analysis of data from the Global Burden of Diseases Study 2015. *Lancet* **389**, 1907–1918 (2017). doi: 10.1016/S0140-6736(17)30505-6; pmid: 28408086
16. India State-Level Disease Burden Initiative Air Pollution Collaborators, The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: The Global Burden of Disease Study 2017. *Lancet Planet. Health* **3**, e26–e39 (2019). doi: 10.1016/S2542-5196(18)30261-4; pmid: 30528905
17. India State-Level Disease Burden Initiative Air Pollution Collaborators, Health and economic impact of air pollution in the states of India: The Global Burden of Disease Study 2019. *Lancet Planet. Health* **5**, e25–e38 (2021). doi: 10.1016/S2542-5196(20)30298-9; pmid: 33357500
18. K. Vohra *et al*., Global mortality from outdoor fine particle pollution generated by fossil fuel combustion: Results from GEOS-Chem. *Environ. Res.* **195**, 110754 (2021). doi: 10.1016/j.envres.2021.110754; pmid: 33577774
19. Centre for Social and Economic Progress, CSEP Electricity and Carbon Tracker (2019); https://carbontracker.in/.
20. Central Electricity Authority, "Executive Summary on Power Sector May 2020" (Government of India, Ministry of Power, 2020).
21. Central Electricity Authority, "Annual Generation Programme 2019-20" (Government of India, Ministry of Power, 2019).
22. R. R. Mohan, N. Dharmala, M. R. Ananthakumar, P. Kumar, A. Bose, "Greenhouse Gas Emission Estimates from the Energy Sector in India at the Subnational Level (Version/edition 2.0)" (GHG Platform India Report, 2019); https://cstep.in/drupal/sites/default/files/2020-05/GHGPI-PhaseIII-Methodology%20Note-Energy-Sep%202019.pdf.
23. S. K. Guttikunda, P. Jawahar, Atmospheric emissions and pollution from the coal-fired thermal power plants in India. *Atmos. Environ.* **92**, 449–460 (2014). doi: 10.1016/j.atmosenv.2014.04.057
24. S. K. Guttikunda, P. Jawahar, Evaluation of Particulate Pollution and Health Impacts from Planned Expansion of Coal-Fired Thermal Power Plants in India Using WRF-CAMx Modeling System. *Aerosol Air Qual. Res.* **18**, 3187–3202 (2018). doi: 10.4209/aaqr.2018.04.0134
25. A. Sehgal, R. Tongia, "Coal Requirement in 2020: A Bottom-up Analysis" (Brookings India, research paper no. 072016-2, 2016); https://www.brookings.edu/wp-content/uploads/2016/09/2016_08_16_coal_future_2020_asrt.pdf.
26. S. K. Sahu, T. Ohara, G. Beig, The role of coal technology in redefining India's climate change agents and other pollutants. *Environ. Res. Lett.* **12**, 105006 (2017). doi: 10.1088/1748-9326/aa814a
27. M. Cropper *et al*., The mortality impacts of current and planned coal-fired power plants in India. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2017936118 (2021). doi: 10.1073/pnas.2017936118; pmid: 33495332
28. W. Peng *et al*., The Critical Role of Policy Enforcement in Achieving Health, Air Quality, and Climate Benefits from India's Clean Electricity Transition. *Environ. Sci. Technol.* **54**, 11720–11731 (2020). doi: 10.1021/acs.est.0c01622; pmid: 32856906
29. M. Gao *et al*., The impact of power generation emissions on ambient $PM_{2.5}$ pollution and human health in China and India. *Environ. Int.* **121**, 250–259 (2018). doi: 10.1016/j.envint.2018.09.015; pmid: 30223201
30. J. Kopas *et al*., Environmental Justice in India: Incidence of Air Pollution from Coal-Fired Power Plants. *Ecol. Econ.* **176**, 106711 (2020). doi: 10.1016/j.ecolecon.2020.106711
31. S. Ramanathan, S. Arora, V. Trivedi, "Coal-Based Power Norms: Where do we stand today?" (Centre for Science and Environment, 2020); https://www.cseindia.org/coal-based-power-norms-coal-based-10125.
32. Ministry of Environment Forest and Climate Change, Government of India, *Gazette of India*, 2015, REGD. NO. D. L.-33004/99.
33. Ministry of Environment Forest and Climate Change, Government of India, *Gazette of India*, 2021, CG-DL-E-01042021-226335.
34. Government of India, "India's Intended Nationally Determined Contribution: Working Towards Climate Justice" (2015); https://www4.unfccc.int/sites/submissions/INDC/Published%20Documents/India/1/INDIA%20INDC%20TO%20UNFCCC.pdf.
35. Power System Operation Corporation Limited, "Security Constrained Economic Dispatch of Inter-state Generating Stations Pan-India: Detailed Feedback Report on Pilot" (2020).
36. H. Safiullah, G. Hug, R. Tongia, Design of load balancing mechanism for Indian electricity markets. *Energy Syst.* **8**, 309–350 (2017). doi: 10.1007/s12667-016-0199-3
37. Ministry of Power, National Power Portal, (2020); https://npp.gov.in/publishedReports.
38. P. Kamboj, R. Tongia, "Indian Railways and Coal: An Unsustainable Interdependency" (Brookings India, 2018); https://www.brookings.edu/wp-content/uploads/2018/07/Railways-and-coal.pdf.
39. S. Srinivasan, N. Roshna, S. Guttikunda, A. Kanudia, S. Saif, J. Asundi, "Benefit Cost Analysis of Emissions Standards for Coal-Based Thermal Power Plants in India" (Center for Study of Science, Technology and Policy, CSTEP report 2018-06, 2018).
40. M. L. Cropper, S. Guttikunda, P. Jawahar, K. Malik, I. Partridge, in *Disease Control Priorities, Third Edition (Volume 7): Injury Prevention and Environmental Health*, C. N. Mock, R. Nugent, O. Kobusingye, K. R. Smith, Eds. (World Bank, 2017), pp. 239–248.
41. M. L. Cropper *et al*., Applying Benefit-Cost Analysis to Air Pollution Control in the Indian Power Sector. *J. Benefit Cost Anal.* **10**, 185–205 (2019). doi: 10.1017/bca.2018.27; pmid: 32968618
42. D. Palchak *et al*., "Greening the Grid: Pathways to Integrate 175 Gigawatts of Renewable Energy Into India's Electric Grid, Vol. I—National Study" (Greening the Grid Program, 2017).
43. D. Palchak *et al*., "Greening the Grid: Pathways to Integrate 175 Gigawatts of Renewable Energy Into India's Electric Grid, Vol. II —Regional Study" (Greening the Grid Program, 2017).
44. D. Palchak, I. Chernyakhovskiy, T. Bowen, V. Narwade, "India 2030 Wind and Solar Integration Study: Interim Report" (National Renewable Energy Laboratory, NREL/TP-6A20-73854, 2019); https://www.nrel.gov/docs/fy19osti/73854.pdf.
45. T. Spencer, N. Rodrigues, R. Pachouri, S. Thakre, G. Renjith, "Renewable Power Pathways: Modelling the Integration of Wind and Solar in India by 2030" (The Energy and Resources Institute, 2020).
46. A. Phadke, N. Abhyankar, R. Deshmukh, "Techno-Economic Assessment of Integrating 175GW of Renewable Energy into the Indian Grid by 2022" (Lawrence Berkeley National Laboratory, 2016).
47. R. Deshmukh, A. Phadke, D. S. Callaway, Least-cost targets and avoided fossil fuel capacity in India's pursuit of renewable energy. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2008128118 (2021). doi: 10.1073/pnas.2008128118; pmid: 33753476
48. T. Lu *et al*., India's potential for integrating solar and on- and offshore wind power into its energy system. *Nat. Commun.* **11**, 4750 (2020). doi: 10.1038/s41467-020-18318-7
49. A. Rose, I. Chernyakhovskiy, D. Palchak, S. Koebrich, M. Joshi, "Least-Cost Pathways for India's Electric Power Sector" (National Renewable Energy Laboratory, NREL/TP-6A20-76153, 2020); https://www.nrel.gov/docs/fy20osti/76153.pdf.
50. N. Abhyankar, S. Deorah, A. Phadke, "Least-Cost Pathway for India's Power System Investments through 2030: A Study Under the Flexible Resources Initiative of the U.S.-India Clean Energy Finance Task Force" (Lawrence Berkeley National Laboratory, 2021); https://eta-publications.lbl.gov/sites/default/files/fri_india_report_v28_wcover.pdf.
51. P. Kumar, R. Banerjee, T. Mishra, A framework for analyzing trade-offs in cost and emissions in power sector. *Energy* **195**, 116949 (2020). doi: 10.1016/j.energy.2020.116949
52. Central Electricity Regulatory Commission, "Discussion Paper on Re-designing Real Time Electricity Markets in India" (2018).
53. Central Electricity Regulatory Commission, "Consultation Paper On Security Constrained Economic Despatch of Inter State Generating Stations Pan India" (2018).
54. Central Electricity Regulatory Commission, "Mechanism for Compensation for Competitively Bid Thermal Generating Stations for Change in Law on Account of Compliance of the Revised Emission Standards of the Ministry of Environment, Forest and Climate Change, Government of India (MoEF&CC)" (2020).
55. "Development of Power Market in India. Phase I: Implementation of Market Based Economic Dispatch (MBED)" (2021); https://powermin.gov.in/sites/default/files/webform/notices/Seeking_comments_on_Discussion_Paper_on_Market_Based_Economic_Dispatch_MBED.pdf.
56. Central Electricity Regulatory Commission, "Discussion Paper on Market Based Economic Dispatch of Electricity : Re-designing of Day-ahead Market (DAM) in India" (2018).
57. C. Oberschelp, S. Pfister, C. E. Raptis, S. Hellweg, Global emission hotspots of coal power generation. *Nat. Sustain.* **2**, 113–121 (2019). doi: 10.1038/s41893-019-0221-6
58. Central Electricity Authority, "Annual Performance of Review of Thermal Power Stations 2014-15" (2015).
59. CMPDI – Coal India Limited, Koyla Grahak Seva (2019); https://elib.cmpdi.co.in/SEVA/index.php.
60. Coal India Limited, "Price Notification" (2018); https://www.coalindia.in/media/documents/Price_Notification_dated_08.01.2018_effective_from_0000_Hrs_of_09.01.2018_09012018.pdf.
61. Government of India, Ministry of Power, Merit Order Despatch of Electricity for Rejuvenation of Income and Transparency (MERIT) (2020); https://meritindia.in/.
62. National Thermal Power Corporation, "Delivered Cost of Gas" (2017).
63. Ministry of Petroleum and Natural Gas, "State/UT-wise Sales Tax Rates Applicable on Crude Oil, Natural Gas and Select Major Petroleum Products As on 1 April, 2018" (Government of India, 2019); https://data.gov.in/resources/stateut-wise-sales-tax-rates-applicable-crude-oil-natural-gas-and-select-major-petroleum.
64. Central Electricity Authority, "Renewable Energy Generation Data" (2018).

65. Power System Operation Corporation Limited, Daily Power Supply Position Reports (2018); https://posoco.in/reports/daily-reports/.
66. Energy Analytics Lab, "Average System Load Profile" (Indian Institute of Technology, 2019).
67. Central Electricity Authority, "Power Allocation from Central Sector" (2020).
68. T. A. Deetjen, I. L. Azevedo, Reduced-Order Dispatch Model for Simulating Marginal Emissions Factors for the United States Power Sector. *Environ. Sci. Technol.* **53**, 10506–10513 (2019). doi: 10.1021/acs.est.9b02500; pmid: 31436968
69. Ministry of Coal, "Provisional Coal Statistics 2017–2018" (2018).
70. D. Singh, R. Tongia, "Need for an Integrated Approach for Coal Power Plants" (CSEP, 2021).
71. U.S. Energy Information Administration, "How much carbon dioxide is produced when different fuels are burned?" (EIA, 2020); https://www.eia.gov/tools/faqs/faq.php?id=73&t=11.
72. International Institute for Sustainable Development, "The Evolution of the Clean Energy Cess on Coal Production in India" (2020); https://www.iisd.org/system/files/publications/stories-g20-india-en.pdf.
73. A. K. Singh, T. B. Kumar, G. Yadav, R. Karna, "Security Constrained Economic Despatch – India: A Rolling Block Implementation Framework," *2019 8th International Conference on Power Systems*, 1–6 (2019).
74. Power Finance Corporation, "Report on Performance of State Power Utilities 2017-18" (2020); https://www.pfcindia.com/DocumentRepository/ckfinder/files/Operations/Performance_Reports_of_State_Power_Utilities/Report_on_Performance_of_State_Power_Utilities_%202017_18.pdf.
75. D. Tong *et al*., Targeted emission reductions from global super-polluting power plant units. *Nat. Sustain.* **1**, 59–68 (2018). doi: 10.1038/s41893-017-0003-y
76. Z. Lu, D. G. Streets, B. de Foy, N. A. Krotkov, Ozone Monitoring Instrument Observations of Interannual Increases in SO_2 Emissions from Indian Coal-Fired Power Plants during 2005–2012. *Environ. Sci. Technol.* **47**, 13993–14000 (2013). doi: 10.1021/es4039648
77. C. Li *et al*., India Is Overtaking China as the World's Largest Emitter of Anthropogenic Sulfur Dioxide. *Sci. Rep.* **7**, 14304 (2017). doi: 10.1038/s41598-017-14639-8; pmid: 29123116
78. A. Soman, K. Ganesan, H. Kaur, "India's Electric Vehicle Transition: Impact on Auto Industry and Building the EV Ecosystem" (Council on Energy, Environment and Water, 2019); https://shaktifoundation.in/wp-content/uploads/2019/10/Indias-Electric-Vehicle-Transition-Report-PDF.pdf.
79. Central Electricity Authority, "CO2 Baseline Database for the Indian Power Sector" (2018).
80. S. Pai, H. Zerriffi, J. Jewell, J. Pathak, Solar has greater techno-economic resource suitability than wind for replacing coal mining jobs. *Environ. Res. Lett.* **15**, 034065 (2020). doi: 10.1088/1748-9326/ab6c6d
81. Central Electricity Authority, "Draft Report on Optimal Generation Capacity Mix for 2029-30" (2019).
82. Central Electricity Authority, "National Electricity Plan - Volume I: Generation" (2018).
83. "The Goods and Services Tax (Compensation to States) Bill, 2017" (PRS Legislative Research, 2017); https://prsindia.org/billtrack/the-goods-and-services-tax-compensation-to-states-bill-2017.
84. S. Pai, H. Zerriffi, A novel dataset for analysing sub-national socioeconomic developments in the Indian coal industry. *IOP SciNotes* **2**, 014001 (2021). doi: 10.1088/2633-1357/abdbbb
85. S. Sengupta *et al*., Reduced-Form Dispatch Model of Indian Power Generation, dataset, *Carnegie Mellon University* (2021); https://doi.org/10.1184/R1/14842224.v1.

ACKNOWLEDGMENTS

The authors thank J. S. Apte for helpful comments in preparing this manuscript. **Funding:** This material is based on work supported by the National Science Foundation Graduate Research Fellowship Program grant no. DGE-1252522 and grant no. DGE-1745016. Any opinions, findings, and conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the National Science Foundation. This publication was developed as part of the Center for Air, Climate, and Energy Solutions (CACES), which was supported under assistance agreement no. R835873 awarded by the US Environmental Protection Agency (EPA). It has not been formally reviewed by the EPA. The views expressed in this document are solely those of authors and do not necessarily reflect those of the EPA. The EPA does not endorse any products or commercial services mentioned in this publication. This work was supported by the Center for Climate and Energy Decision Making (SES-1463492) through a cooperative agreement between the National Science Foundation and Carnegie Mellon University. Funding for this research was supported by the National Security Education Program's Boren Fellowship. **Author contributions:** S.S. contributed to the conceptualization, gathered data, created the model, analyzed results, and wrote the paper. P.J.A. contributed to the conceptualization, analyzed results, and wrote the paper. T.A.D. contributed to the conceptualization. S.D. gathered data. P.K. gathered data. R.T. contributed to conceptualization, analyzed results, and wrote the paper. I.M.L.A. contributed to the conceptualization, analyzed results, and wrote the paper. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Model source code, input data, and modeling results are available online (*85*). **License information:** Copyright © 2022 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abh1484
Supplementary Text
Figs. S1 to S31
Tables S1 to S4
References (*86*, *87*)

Submitted 1 June 2021; accepted 23 September 2022
10.1126/science.abh1484

RESEARCH ARTICLE SUMMARY

FISHERIES

Seventy years of tunas, billfishes, and sharks as sentinels of global ocean health

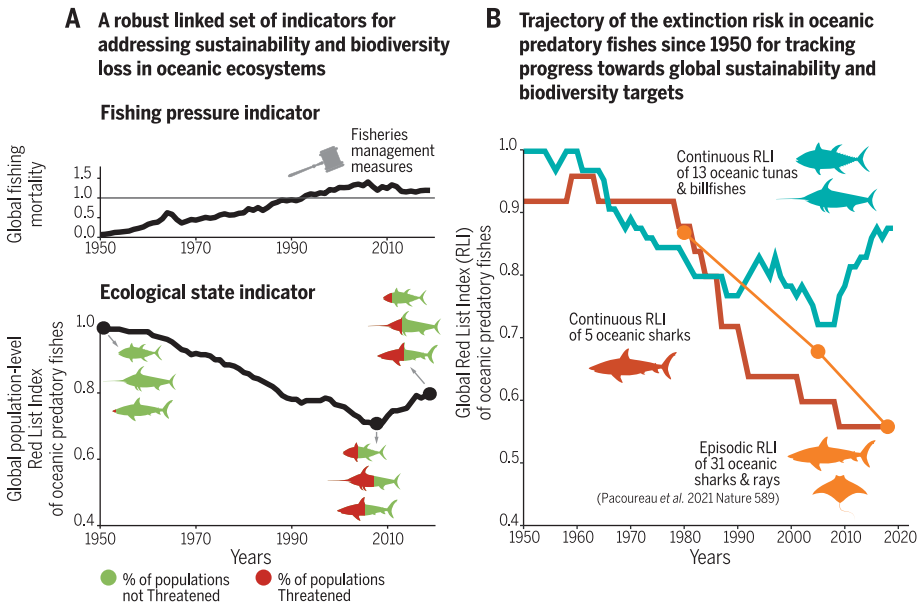
Maria José Juan-Jordá*, Hilario Murua, Haritz Arrizabalaga, Gorka Merino, Nathan Pacoureau, Nicholas K. Dulvy

INTRODUCTION: Recent biodiversity assessments show unprecedented loss of species, ecosystems, and genetic diversity on land but it remains unclear how widespread such patterns may be in the oceans. There is an urgent need to develop surveillance indicators to track the health of ecosystems in the marine realm, including changing extinction risk of marine species. These will allow evaluation of progress toward achieving global goals and commitments established by the Convention of Biological Diversity (CBD) and Sustainable Development Goals (SDGs) to halt and reverse marine biodiversity loss.

RATIONALE: Highly monitored oceanic fisheries comprising iconic predatory tunas, billfishes, and sharks yield an opportunity to support the development of linked sets of pressure and ecological state indicators capable of measuring progress toward global biodiversity and sustainability targets. We derived a continuous Red List Index (RLI) based on International Union for Conservation of Nature (IUCN) Red List categories and criteria for tracking yearly changes in extinction risk of oceanic tunas, billfishes, and sharks over the past 70 years to assess the health of oceanic biodiversity. Furthermore, by assessing the sensitivity and responsiveness of the RLI (state indicator) to fishing mortality (pressure indicator) and assessing the alignment between the most recent Red List status and fishery exploitation status of tunas, billfishes, and shark populations, we offer decision-makers a robust set of linked pressure-state indicators for tracking biodiversity loss and recovery in oceanic ecosystems.

RESULTS: We find that since 1950, the global extinction risk of oceanic predatory fishes has continuously worsened as a result of rising and excessive fishing pressure, up until the late 2000s when management actions reduced fishing mortality, allowing for recovery of tunas and billfishes. However, sharks remain undermanaged and their extinction risk continues to rise. Our findings reveal a core problem and ongoing challenge in the management of oceanic multigear and multispecies fisheries. Whereas target species are increasingly sustainably managed to ensure maximum yields, the functionally important shark species being captured incidentally by the same fisheries continue to decline as a result of insufficient management actions. Furthermore, our study also connects annual changes in global extinction risk with changes in fishing mortality over the last 70 years, demonstrating how the global RLI trajectory of oceanic predatory fishes is highly sensitive and responsive to fishing mortality.

CONCLUSION: Although halting biodiversity loss by rebuilding highly valuable commercial tuna and billfish species has been achieved, the next challenge is to halt declines in shark species by setting clear biodiversity goals and targets as well as implementing science-based conservation and fishery management measures and international trade regulations. Unless an effective mitigation hierarchy of management actions to reduce shark mortality is urgently implemented (and adapted to the complexity of each fishery and shark species), their risk of extinction will continue to increase. Furthermore, we demonstrate a high alignment and complementarity between the current population-level Red List status and fishery exploitation status of tunas, billfishes, and sharks, when applied at the same scale. Although we do not propose that the RLI be used to manage fish populations, this strong alignment eliminates any technical barrier for use of the RLI by policy-makers for tracking CBD and SDG targets. ■



Global Red List Index (RLI) of oceanic predatory fishes for tracking progress toward global biodiversity and sustainability targets. (**A**) The global population-level RLI (state indicator) closely tracks changes in fishing mortality (pressure indicator) for 52 oceanic tuna, billfish, and shark populations over the last 70 years, thus providing decision-makers with a linked set of pressure-state indicators for tracking the health of oceanic biodiversity. The population-level RLI was reversed in 2008 following a reduction in fishing mortality after implementation of fisheries management measures in tuna regional fisheries management organizations. The horizontal gray line denotes F/F_{MSY} =1, F_{MSY} being fishing mortality (F) which produces the maximum sustainable yield (MSY). (**B**) Global continuous species-level RLI of tunas, billfishes, and oceanic sharks (seven, six, and five species, respectively) tracking yearly changes in extinction risk over 70 years and the global episodic RLI of oceanic sharks and rays (21 and 10 species, respectively) estimated in 1980, 2005, and 2018. An RLI value of 1 indicates that a given taxa qualifies as least concern (that is, not expected to become extinct in the near future), whereas an RLI value of zero indicates that all taxa have gone extinct.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: mjuan@azti.es
Cite this article as M. J. Juan-Jordá *et al.*, Science 378, eabj0211 (2022). DOI: 10.1126/science.abj0211

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abj0211

RESEARCH ARTICLE

FISHERIES

Seventy years of tunas, billfishes, and sharks as sentinels of global ocean health

Maria José Juan-Jordá[1]*, Hilario Murua[2], Haritz Arrizabalaga[1], Gorka Merino[1], Nathan Pacoureau[3], Nicholas K. Dulvy[3]

Fishing activity is closely monitored to an increasing degree, but its effects on biodiversity have not received such attention. Using iconic and well-studied fish species such as tunas, billfishes, and sharks, we calculate a continuous Red List Index of yearly changes in extinction risk over 70 years to track progress toward global sustainability and biodiversity targets. We show that this well-established biodiversity indicator is highly sensitive and responsive to fishing mortality. After ~58 years of increasing risk of extinction, effective fisheries management has shifted the biodiversity loss curve for tunas and billfishes, whereas the curve continues to worsen for sharks, which are highly undermanaged. While populations of highly valuable commercial species are being rebuilt, the next management challenge is to halt and reverse the harm afflicted by these same fisheries to broad oceanic biodiversity.

Recent global biodiversity assessments show unprecedented human-driven declines in abundance of wild species, compromising the integrity and functioning of ecosystems on Earth (*1*, *2*). However, the scale of damage upon oceanic ecosystems remains unclear. Fishing activity is increasingly monitored by satellites (*3*) and fishery statistics (*4*), but its effects on ocean biodiversity are not similarly tracked. The Convention of Biological Diversity (CBD) and the Sustainable Development Goals (SDGs) of the 2030 Agenda for Sustainable Development together established a framework of agreed-upon targets and actions for governments to reduce the current rate of biodiversity loss at the global, regional, and national scale. This requires linked sets of pressure and ecological state indicators capable of measuring progress toward achieving global marine biodiversity and sustainability targets (*5*, *6*).

Several major oceanic predatory fishes—tunas, billfishes, and sharks—exhibit three features that make them strong candidates for assessment of the trajectory of oceanic biodiversity (Fig. 1, fig. S1, and table S1). First, they are among the largest (100 to 500 cm) megafaunal predators and most functionally unique species in pelagic ecosystems, and they play a critical role in regulating the structure, function, and stability of oceanic ecosystems (*7*). Second, they exhibit differential resilience to overfishing and span a range of fisheries categories from economically valuable target species to ecologically important incidental catch (*8*, *9*). Third, they are routinely monitored and assessed by the five tuna regional fisheries management organizations (tuna RFMOs) with the mandate to conserve and manage transboundary large migratory fish species (table S2 and fig. S2). Time series of biomass and fishing mortality rates derived from fish stock assessments are available for 52 populations of 18 species, encompassing 60% of oceanic predatory fish diversity (Fig. 1 and figs. S3 to S6). This data richness enables the development of linked sets of pressure and ecological state indicators capable of tracking global targets.

The International Union for the Conservation of Nature (IUCN) Red List Index (RLI) is a well-established ecological state indicator adopted as one of the official UN SDG and CBD indicators (*5*). The RLI is based on the IUCN Red List categories and criteria, which uses one of five quantitative criteria (A to E) to classify species into one of eight categories of extinction risk: extinct (EX), extinct in the wild (EW), critically endangered (CR), endangered (EN), vulnerable (VU), near threatened (NT), least concern (LC), and data deficient (DD) (*10*, *11*). The RLI shows trends in the overall extinction risk for a group of species by measuring how the number of species in each Red List category changes over time scaled from 1 (all species LC) to 0 (all species EX). The RLI has already been estimated from the episodic application of the IUCN Red List categories and criteria to the world's birds, mammals, amphibians, corals, cycads, and oceanic sharks and rays (*2*, *9*) by Red List Authorities and Specialist Groups of the IUCN Species Survival Commission. These episodic Red List assessments occur every 4 to 10 years, thus far yielding time series of 2 to 4 data points spanning up to four decades.

We first derive a novel continuous year-on-year RLI using a Bayesian framework to model population time series and estimate probabilistic extinction risk applying the IUCN Red List A criterion (fig. S7) (*12*, *13*). Then, we develop a global continuous RLI for 18 oceanic predatory fishes of tunas, billfishes, and sharks from 1950 to 2019 to assess the state of oceanic biodiversity. Finally, we assess the sensitivity and responsiveness of the RLI trajectory to fishing pressure, providing decision makers with an integrated linked set of pressure-state indicators for tracking biodiversity change.

We illustrate our six-step method to estimate extinction risk with its application to the Southern Bluefin Tuna (*Thunnus maccoyii*; Fig. 2 and figs. S7 to S9) (*14*). Criteria A classifies extinction risk based on exceeding a threshold of population decline over the greater part of 10 years or three generation lengths (GL). First, we defined the GL of the given species (12 years) and extracted abundance time series from the most recent fish stock assessment (Fig. 2A, fig. S5, and table S3). Second, we calculated the total percent change in population biomass over three GL by estimating the average annual rate of population change over the

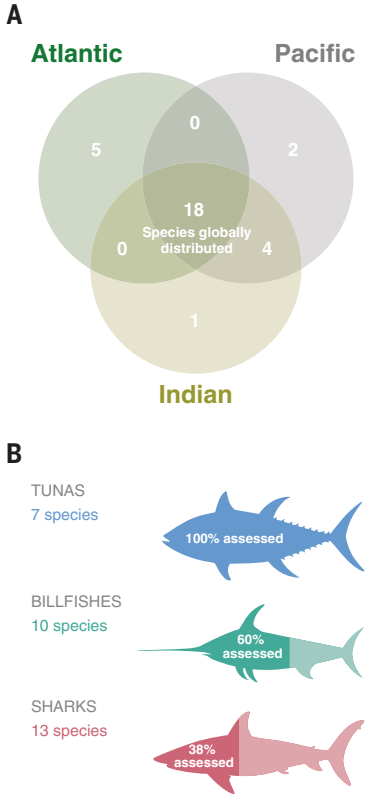


Fig. 1. Oceanic predatory fishes of the world. (**A**) Total number of oceanic tunas, billfishes, and sharks distributed globally and by ocean (table S1). (**B**) Proportion of species with at least one population assessed with fish stock assessment models by major taxa (table S2).

[1]AZTI, Marine Research, Basque Research and Technology Alliance (BRTA). Herrera Kaia, Portualdea z/g, 20110 Pasaia, Gipuzkoa, Spain. [2]International Seafood Sustainability Foundation, Pittsburgh, PA, USA. [3]Earth to Ocean Research Group, Department of Biological Sciences, Simon Fraser University, Burnaby, British Columbia, Canada.
*Corresponding author. Email: mjuan@azti.es

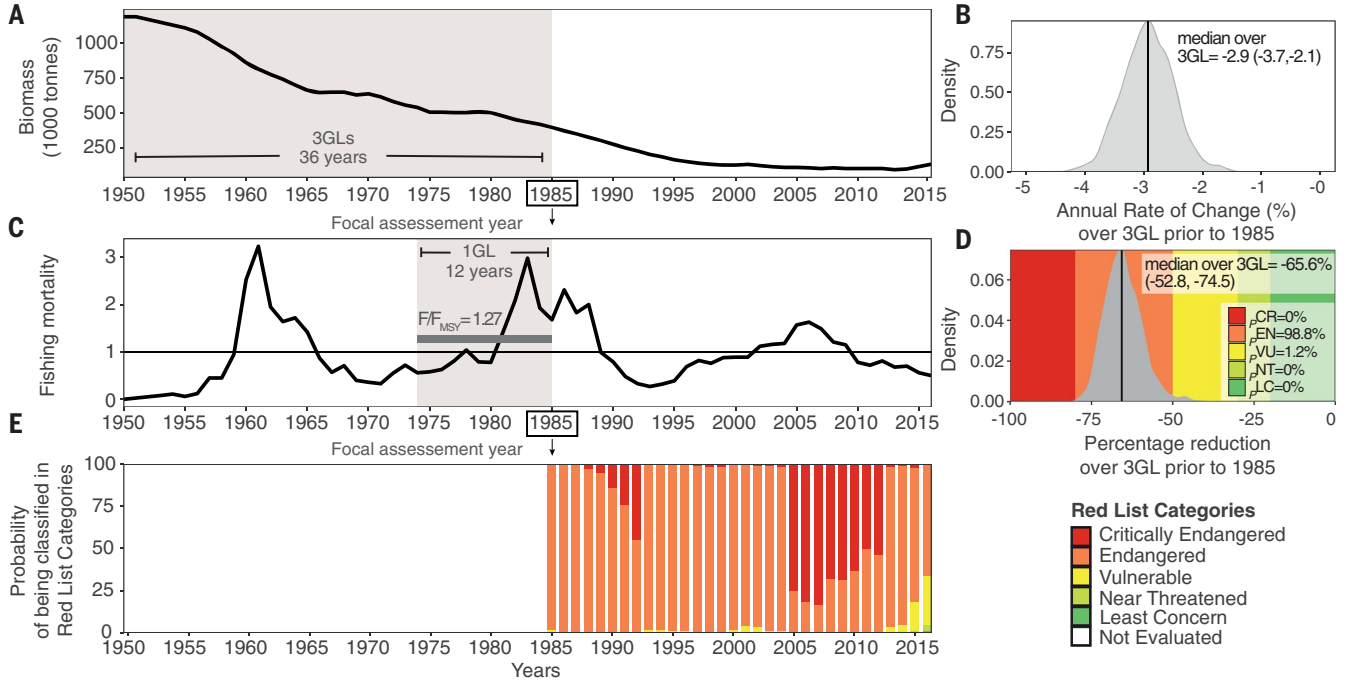


Fig. 2. Illustrative example of a continuous Red List assessment using Criterion A for Southern Bluefin Tuna from 1985 to 2016. (**A**) Time series of biomass from the latest fish stock assessment (table S2). The shaded rectangle shows the three GL window used to estimate the Red List category for 1985. (**B**) Posterior probability distribution and median (vertical black line) of the estimated average annual rate of change (percent) in population size over the previous three GL in 1985. (**C**) Time series of fishing mortality rate relative to F_{MSY}. The shaded rectangle shows the average fishing mortality over a one GL window before 1985, showing that the species is not being sustainably managed—hence application of the A2 thresholds. (**D**) Posterior median (vertical black line) and probability distribution of the estimated total reduction over three GL in 1985. The posterior probability is overlaid on the Red List category A2 thresholds. (**E**) The probability of being classified in the Red List categories in 1985 and at each subsequent year between 1985 and 2016.

three GL window using an intercept-only hierarchical Bayesian model (*14*). These models allow for nonlinearity in population trends and account for the hierarchical structure of the data as some species trends are based on multiple population estimates from multiple fish stock assessment models (fig. S5). We estimated that by 1985, Southern Bluefin Tuna had a median population reduction of 65.6% [95% credible interval (CI) 52.8, 74.5], equivalent to an annual rate of change of −2.9% (CI −3.7, −2.1) (Fig. 2B). Third, we classified status using either A1 thresholds, when the species is sustainably managed worldwide (i.e., the causes of decline are reversible, and understood, and have ceased) in at least 90% of its range, or A2 thresholds otherwise. Specifically, A1 thresholds for population reduction (VU = 50 to 69%, EN = 70 to 89%, and CR ≥ 90%) are applied to sustainably managed species. In operational terms, a fish species is considered sustainably managed when the average fishing mortality (F) on the species is below the fishing mortality corresponding to the maximum sustainable yield (MSY) ($F/F_{MSY} \leq 1$) for the previous one GL in at least 90% of its range in accordance with IUCN guidelines (*15*). Otherwise, for unsustainably managed species ($F/F_{MSY} > 1$), the A2 thresholds (VU = 30 to 49%, EN = 50 to 79%, and CR ≥ 80%) are applied. In our illustrative example, the Southern Bluefin Tuna was not being sustainably managed ($F/F_{MSY} = 1.27$) in 1985 based on the average fishing mortality over the one GL window before 1985, hence application of the A2 threshold (Fig. 2C). Fourth, we assigned Red List category probabilities because the Bayesian estimation framework allows us to propagate the uncertainty in population reductions into probabilistic classifications for each of the Red List categories (*12*). Based on a population reduction value of 65.6%, Southern Bluefin Tuna was classified as EN (probability $_pEN=98.7\%$ and $_pVU=1.3\%$) in 1985 (Fig. 2D). The fifth step consisted of a year-on-year estimation of Red List status for the entire time series, which reveals how Southern Bluefin Tuna became increasingly threatened over time to the point where in 2005 it was classified as CR ($_pCR=76\%$ and $_pEN=24\%$; Fig. 2E). As fishing mortality was reduced from 2006 onward, the biomass of Southern Bluefin Tuna stabilized at low levels and has recently started to increase—this is closely tracked by a reduction in extinction risk in the most recent years ($_pCR=0\%$, $_pEN=66\%$, $_pVU=29\%$, $_pNT=4\%$, and $_pLC=1\%$ in 2016; Fig. 2E). This is a case for which we have one population representing the whole species. An example of a species composed of multiple populations and how they are combined to the species level is available in the supplementary materials (fig. S10 and table S4). Last, we aggregated the Red List status hierarchically across populations (fig. S8) and then species (fig. S9) to derive the global RLI of oceanic predatory fishes (*14*).

Since 1950, the global RLI trajectory of oceanic predatory fishes worsened by ~27% (95% CI 24.4, 31.1) reflecting the increasing extinction risk of the whole assemblage until recovery became apparent in 2008 (Fig. 3A). In that year, 10 species were classified as threatened, with Southern Bluefin Tuna and Oceanic Whitetip Shark (*Carcharhinus longimanus*) classified as CR; Blue Marlin (*Makaira nigricans*), Silky Shark (*Carcharhinus falciformis*), Porbeagle Shark (*Lamna nasus*), and Swordfish (*Xiphias gladius*) as EN; and Bigeye Tuna (*Thunnus obesus*), Yellowfin Tuna (*Thunnus albacares*), Pacific Bluefin Tuna (*Thunnus orientalis*), and Striped Marlin (*Kajikia audax*) as VU (fig. S9). The most recent recovery of the RLI since 2008 reflects improvement (from CR to VU) of Southern Bluefin Tuna, and improvement of five species into NT and LC [Yellowfin Tuna, Swordfish, Blue Marlin, Striped Marlin, and Black Marlin (*Makaira indica*); fig. S9]. However, the RLI trajectory varies among major taxa (Fig. 3, B and C). For tunas, the RLI started to improve in the 1990s and end of the 2000s

whereas the RLI for billfishes deteriorated until the early 2000s, improving only during the past decade. However, the RLI of sharks has worsened continuously. Our continuous RLI is robust to the choice of different time windows for calculating fishing mortality metrics and population range–based scenarios to determine whether a species is being sustainably managed throughout its entire range (figs. S11 to S13) (*14*).

To understand how changes in population-level fishing mortality underlie the RLI, we derived a global population-level RLI for the 52 assessed populations (Fig. 4 and fig. S14). The 58-year decline and recent recovery in the population-level RLI of oceanic predatory fishes closely tracks the historical trend of fisheries development and implementation of fisheries management in these species. Since the 1950s, global average fishing mortality has been increasing, exceeding sustainable levels in 1993 and then peaking in 2006 (Fig. 4A). Over this same period, the average biomass of oceanic predatory fishes declined

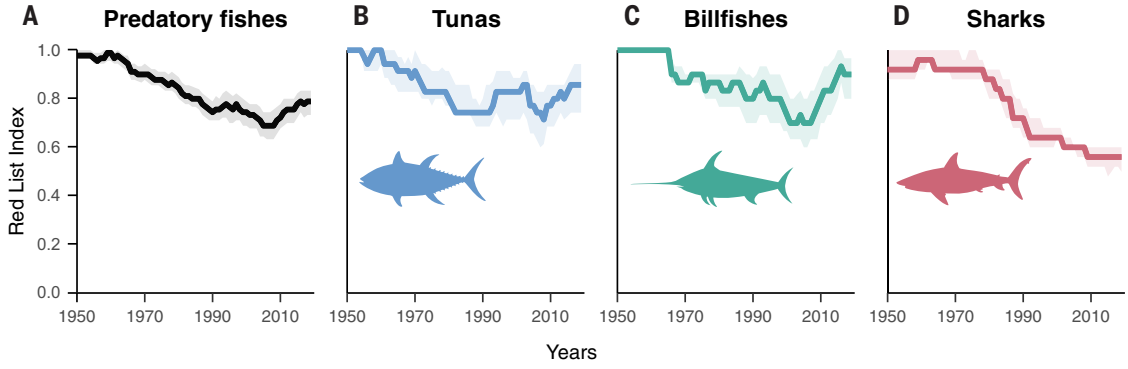


Fig. 3. RLI of oceanic predatory fishes. (**A**) The global RLI includes 18 species of oceanic tunas, billfishes, and sharks and is disaggregated by major taxon: (**B**) tunas, (**C**) billfishes, and (**D**) sharks. The solid line denotes the median and the shaded polygons denote the 95% CI. An RLI value of 1.0 indicates that all species qualify as Least Concern (that is, not expected to become extinct in the near future) whereas an RLI value of 0 indicates that all species have gone extinct.

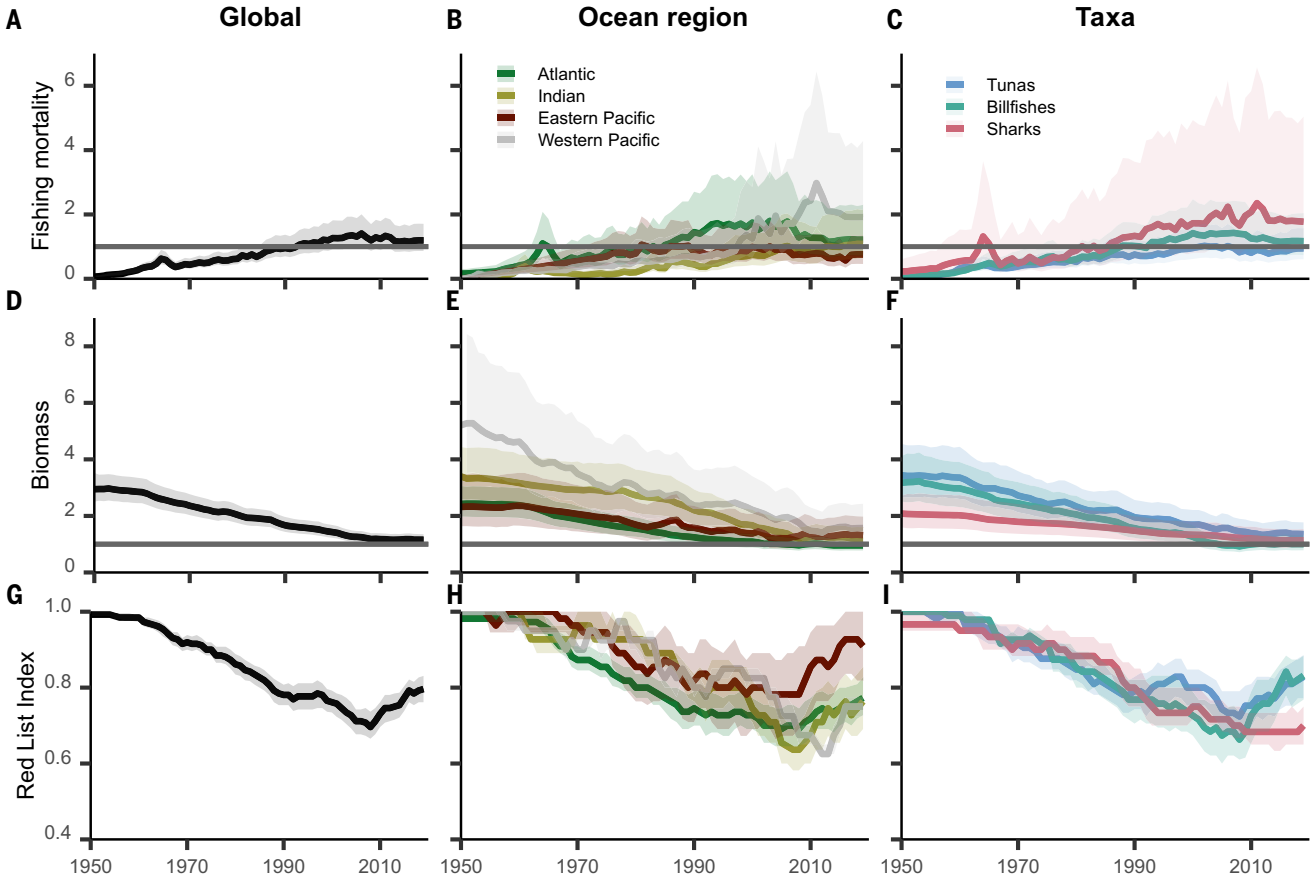


Fig. 4. Trends in overall fishing mortality and their impact on population biomass and population-level RLI trajectory of oceanic predatory fishes. (**A**) Global average fishing mortality rates relative to F_{MSY} and is disaggregated by major ocean regions (**B**) and taxon (**C**). (**D**) Global average biomass relative to B_{MSY} and is disaggregated by major ocean regions (**E**) and taxon (**F**). (**G**) Global population-level RLI and is disaggregated by major ocean regions (**H**) and taxon (**I**). The solid line denotes the median and the shaded polygons the 95% CIs. The horizontal gray lines denote the F_{MSY} and B_{MSY}. Interpretation of RLI values can be found in Fig. 3.

and then approached the MSY (B_{MSY}; Fig. 4D). Consequently, the population-level RLI of oceanic fishes worsened steadily since the 1950s, reaching its lowest value in 2008, 2 years after the maximum value of fishing mortality (Fig. 4G). When fishing mortality started to decrease after 2006, the population-level RLI reversed shortly after, reflecting the reclassification of many populations into less threatened categories (fig. S8).

The extent and timing of management measures implemented by tuna RFMOs differ markedly among ocean regions and taxa, influencing overall fishing mortality, biomass, and population-level RLI trajectories (Fig. 4). Regionally, the RLI trajectories track the historical increase in fishing mortality following the development of industrial tuna fisheries, which began first in the Atlantic and eastern Pacific before expanding to the Indian and western Pacific oceans during the 1980s (Fig. 4B). The lowest RLI values observed in the Indian and western Pacific around the 2010s (Fig. 4H) were due to the steep decline in biomass (Fig. 4E) resulting from the rapid increase in fishing mortality. We also find that the different timing in the stabilization pattern of overall biomass levels around the management target of MSY in the four ocean regions has resulted in the observed region-level reductions in extinction risk. The RLI has been reversed in all oceans through reductions in fishing mortality (Fig. 4H). When examining population-level RLI trajectories by major taxon, we confirm that the declining RLI trajectory has not only been halted but also reversed for tunas and billfishes (Fig. 4I). We attribute these recoveries to a reduction in overall fishing mortality (Fig. 4C) and hence the recovery of biomass toward sustainable levels (Fig. 4F) following effective management measures. However, we caution that the threatened status of some tunas and billfishes (e.g., Indian Ocean Yellowfin and Atlantic Bigeye Tuna) require strengthened management measures (fig. S8). Historically, sharks have been the incidental catch of these tuna and billfish fisheries and have declined steeply (Fig. 4, F and I) as fishing mortality is twice that of the sustainable level (Fig. 4C). Despite increasing scientific evidence and public concern, undermanaged populations of oceanic sharks continue to worsen along a path of increasing extinction risk (Fig. 4I).

By next demonstrating the correlation between fishing mortality and the RLI and evaluating the alignment between fishery exploitation status and Red List status, we offer decision-making tools for tracking and tackling biodiversity loss in oceanic ecosystems thus supporting UN CBD and SDG processes (*6*). First, we assess the sensitivity and responsiveness of the RLI trajectory to fishing pressure using a prewhitened cross-correlation analysis for removing the autocorrelation and trends in the time series. We show that the RLI closely tracks changes in fishing mortality and find a significant negative cross-correlation between fishing mortality and the RLI (Fig. 5A), suggesting that the RLI is sensitive (sensitivity = −0.34) and responsive to fishing mortality (with a significant time lag of only 2 years) (*14*) (figs. S15 to S17). The pressure-state relationship is reversible and symmetric, with the RLI recovering as fishing mortality decreases, tracking back (green points) along the same path as the decline trajectory (red points, Fig. 5A). Second, we assessed alignment by comparing the fishery exploitation status [whether populations are considered overfished ($B < B_{MSY}$) or not ($B \geq B_{MSY}$), derived from the latest fish stock assessments and the corresponding population-scale Red List status (*14*) (table S5). We find that the fishery exploitation status and Red List status are aligned in 76.6% of the assessments (true positives and negatives; Fig. 5B and table S6). Therefore, a sustainable fishery will have low extinction risk, and conversely in an unsustainable fishery, an overfished population will likely have a higher extinction risk. However, some overfished populations were categorized in the low-risk category of LC by the Red List assessment (a "miss" of 12.8%; Fig. 5B), as they may not be considered threatened when their abundance declines have been stabilized at low levels and the causes of decline are understood and have ceased. Furthermore, there were few "false alarms" (10.6%) in which the Red List criteria classified a population as threatened although it was not being estimated as overfished, offering an early warning for those populations with relatively large biomass declining rapidly toward target levels (fig. S5). These false alarms are transient and disappear if populations are stabilized at target levels. Altogether, this harmony in criteria sets is highly consistent with all other modeling and meta analyses comparing the Red List status with fishery exploitation status over a wide range of marine fishes (*16*–*18*) and provides further evidence of alignment among both classification systems when applied at the same scale. Although we do not propose that the RLI be used to manage populations, there should be no concerns that a threatened listing is inconsistent with fishery management advice as these mismatches can often be understood and explained. Hence, our findings of strong alignment demonstrate that both criteria sets are complementary and eliminate any technical

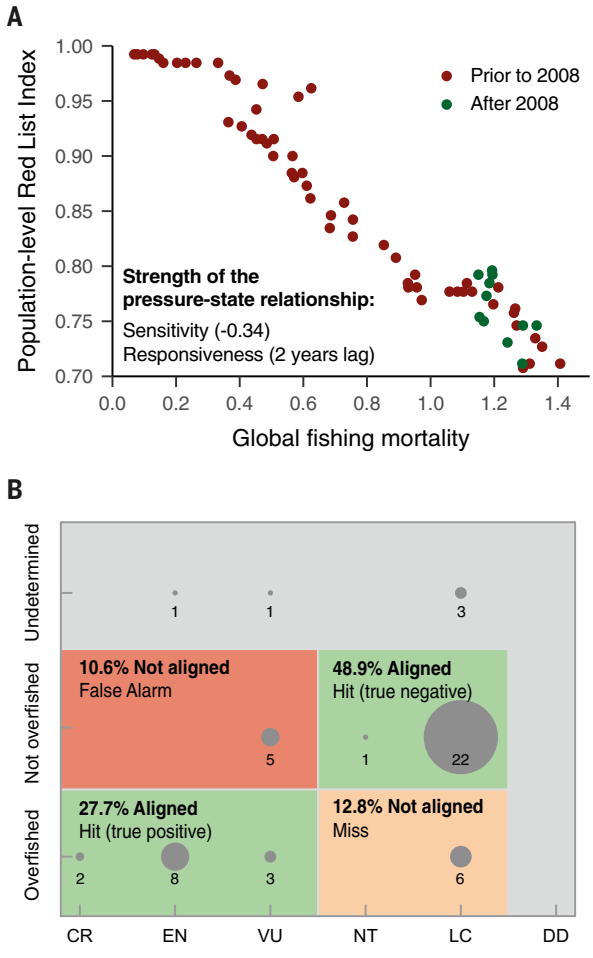


Fig. 5. Effects of fishing mortality on the state of oceanic predatory fishes. (**A**) Prewhitened cross-correlation between global average annual fishing mortality and population-level RLI for the assessment of sensitivity and responsiveness of the RLI to fishing. (**B**) Alignment between the population-level Red List status in relation to fishery exploitation status. Current fishery exploitation status, whether the population is considered overfished ($B_{current} < B_{MSY}$) or not ($B_{current} > B_{MSY}$), derived from the most current fishery assessments (*y* axis) and the Red List status for the same assessment year (*x* axis). Circle size is proportional to the number of populations classified in each category.

barrier for use of the RLI by policy-makers for tracking CBD and SDG targets.

Our continuous RLI of oceanic predatory fishes advances and complements episodic RLI as calculated for other animal and plant groups (Fig. 6) as it allows for tracking of status and trends in extinction risk on much finer time scales. In a half a century, industrial fisheries have reduced oceanic pelagic biodiversity to levels similar to those brought about for other terrestrial taxa over the course of centuries (*19*). The initial warnings now seem timely and appropriate given how rapidly the RLI of oceanic tunas, billfishes, and some sharks have recovered to levels more typical of terrestrial vertebrates. This provides evidence that decisive action by fisheries agencies can recover exploited fishes, but we have yet to take similarly decisive action for sharks. Furthermore, our continuous RLI trajectories for tunas (Fig. 3B) and sharks (Fig. 3D) are highly consistent with the recently published episodic IUCN Red List assessments for oceanic tunas and sharks (*8*, *9*, *20*), showing that the RLI for oceanic tunas has recovered between 2011 and 2021 and that the global extinction risk for sharks continues to worsen. For data-rich taxa, both the episodic and continuous RLI are highly aligned because both Red List assessments are driven by the same data, though we note that the episodic formal IUCN Red List assessments process has scope to diverge as it considers other criteria (B to E), threats, use and

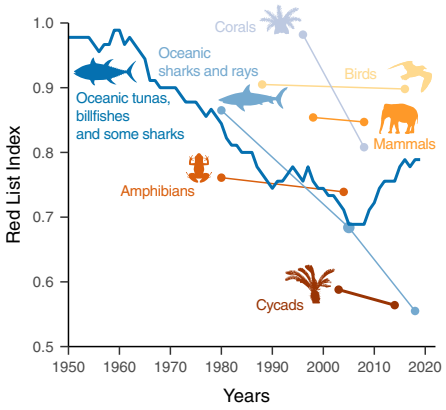


Fig. 6. Decline and recovery of the RLI of oceanic predatory fishes in the context of increasing risk of extinctions in major taxa groups. Our species-level RLI of oceanic tunas (n = 7 species), billfishes (n = 6), and sharks (n = 5) adds to the already monitored episodic RLI trajectories of marine taxa groups (illustrated with tones of blue): oceanic sharks and rays (n = 31; 5 of these species are included in our continuous RLI for sharks), and corals (n = 704). Terrestrial taxa groups are illustrated with earthy tones: mammals (n = 4556), birds (n = 9869), amphibians (n = 4355), and cycads (n = 307) (*2*, *9*).

trade, and conservation actions to categorize species in addition to the population reduction analysis used here. Finally, our continuous RLI could be applied to other marine fishes and any other taxa with time series of population data, which would increase the temporal and spatial resolution of both global and regional Red List and RLI assessments. We reaffirm the need to expand the representation of marine species on the Red List to monitor marine biodiversity because most marine taxa remain unassessed (*21*).

Our study connects annual changes in fishing mortality and extinction risk globally over the past 70 years for oceanic tunas, billfishes, and sharks and reveals how effective management for highly valuable commercial species of tunas and billfishes has reversed the biodiversity loss curve while the extinction risk of undermanaged sharks continues to increase. Our vignette of oceanic predator fisheries reveals the biggest challenge of global multigear and multispecies fisheries management, as target species are increasingly being brought to sustainable levels to ensure maximum yield. However, the shark species incidentally captured by the same fisheries continue to decline to the point where there is increasing risk of biodiversity loss due to insufficient management actions (*22*, *23*). Driven by policy complexity, insufficient data and monitoring, socioeconomic concerns, and lack of political action, oceanic sharks remain undermanaged and a lower priority in tuna RFMOs despite repeated and increasingly intense warnings based on their high intrinsic sensitivity to overfishing, increasing catches, and the high international trade value of their meat and fins (*9*, *24*). To date, conservation and management measures in tuna RFMOs for sharks remain largely focused on mitigating the effects of fishing on incidental catches through gear modification (e.g., banning shark leaders), safe handling and release practices (e.g., devil rays caught in purse seines), prohibition of retention (e.g., thresher and hammerheads sharks), and establishing requirements for data reporting to support their assessments (*24*). However, there seems to be high resistance to any measure that might meaningfully curb fishing mortality for sharks. Unless an effective mitigation hierarchy of management actions to reduce shark mortality—including international trade regulation—are urgently implemented and adapted to the complexity of each fishery and shark species, their trajectories will continue worsening in the future (*25*). We show that reversing the curve of oceanic biodiversity loss is possible in the case where fishery sustainability goals and effective management measures are implemented, even in the challenging context of international fisheries management. Defining new priorities and setting clear biodiversity goals and targets to halt and reverse broad oceanic biodiversity loss remains the next management challenge to achieve progress for both people and oceanic biodiversity.

Materials and Methods

Compilation of population data from fish stock assessments

We compiled the most recent (as of June 2020) fish stock assessments for 52 populations (18 species) of tunas, billfishes, and sharks from the five tuna RFMOs (fig. S2 and table S2) (*14*). For each fish stock assessment, we extracted the following: (a) the estimated time series of biomass or time series of biomass relative to the biomass that produces the MSY [B/B_{MSY}] (fig. S5), (b) the estimated time series of fishing mortality relative to the fishing mortality that produces the MSY [F/F_{MSY}] (fig. S6), and (c) the standard biological reference points used to determine population status, generally the current adult biomass relative to the adult biomass producing MSY ($B_{current}/B_{MSY}$) and current fishing mortality rate relative to the fishing mortality that maintains MSY ($F_{current}/F_{MSY}$) (table S2). This data was extracted from the assessment models (and model runs) used to determine population status and provide management advice by the Scientific Committees of each of the tuna RFMOs (*14*).

Compilation and estimation of generation lengths

We also collated the GL for each species (and populations) of tunas, billfishes, and sharks from the published literature or as approved for use by the IUCN Tuna and Billfish Specialist Group or the IUCN Shark Specialist Group (table S3). In some cases we also estimated GL for populations using age-structured life tables (*14*).

Estimation of Red List status

We applied the IUCN Red List categories and criteria to calculate the extinction risk for 18 species of tunas, billfishes, and sharks (fig. S7) (*14*). All species of oceanic tunas, billfishes, and sharks were assessed under IUCN Red List Criterion A "population reduction." Criterion A was applied to both the taxonomic unit of population and the taxonomic unit of species, to assign a Red List category to each population and species of tunas, billfishes, and sharks between 1950 and 2019 (figs. S8 and S9). For each species and population, we estimated the total percent change in biomass within the past three GL yearly between 1950 and 2019, and then we assigned a Red List category using Criterion A1 or A2 thresholds, depending on whether the species/population was being effectively and sustainably managed. A fish species/population is considered sustainably managed when the average fishing mortality

(F) on the species or population is below the fishing mortality corresponding to the maximum sustainable yield (MSY; $F/F_{MSY} \leq 1$) for the previous one GL in at least 90% of its range according to IUCN guidelines (*15*). When estimating the total percent change for species with multiple populations, we weighted the estimated total percent change in biomass of each population by their MSY to account for the contribution of different population sizes to the global species (table S2). We calculated the total percent change in population biomass over the past three GL by estimating the average annual rate of population change over the three-GL window using an intercept-only hierarchical Bayesian model (*14*). At the end, we were able to assign Red List categories to the taxonomic unit of population (fig. S8) and the taxonomic unit of species (fig. S9) annually between 1950 to 2019 (fig. S10 shows two illustrative examples of Red List status calculations). Because of the Bayesian estimation framework, we assigned Red List category probabilities allowing us to propagate the uncertainty in population reductions into probabilistic classifications for each of the Red List categories. The application of A1 or A2 thresholds for assigning the most likely Red List category requires to determine on an annual basis whether a population and species is being sustainably managed. We conducted two different sensitivity analyses for evaluating the impact of calculating in different ways whether a population and species is being sustainably managed on the determination of extinction risk (figs. S11, S12, and S13) (*14*).

Estimation of RLI

We calculated a continuous RLI for oceanic predatory fishes between 1950 and 2019 using the estimated extinction risk of the 18 species of tunas, billfishes, and sharks (figs. S7 and S14) (*14*). We also disaggregated the global species-level RLI by major taxon (tunas, billfishes, sharks). Traditionally, the RLI is calculated from the episodic application of the IUCN Red List categories and criteria to species groups, and this usually occurs episodically involving Red List Authorities and Specialist Groups of the IUCN Species Survival Commission. Instead, here, we calculated a continuous RLI using a Bayesian framework by classifying species into the extinction risk probabilistic categories using time series analyses of population data derived from fish stock assessments (*14*). The Bayesian estimation framework improved the characterization of uncertainty in the RLI, which facilitates the communication of uncertainty and probabilistic statements to conservation practitioners (*12*).

We also calculated a global population-level RLI in order to examine the effects of global fishing pressure (here expressed as fishing mortalities), which is monitored at the level of population, on the population-level RLI trajectories. We calculated the global population-level RLI using the Red List status of the 52 populations of tunas, billfishes, and sharks as the basic unit of assessment (instead of using species as the unit of assessment) assigning equal weighting to all populations (fig. S7 and S14) (*14*).

Estimation of the overall trajectories of biomass and fishing mortalities

We calculated the global overall trajectory of biomass and fishing mortality across the 52 populations of oceanic predatory fishes from 1950 to 2019 by fitting a Bayesian generalized linear model where the fishing mortality or biomass values were treated as the response variable with a Gamma likelihood and an identity link function, and the years were the fixed predictors (treated as factors) (*14*). In this way, the estimated average biomass or fishing mortality had balanced data as each year had the same number of populations and each population weighted equally.

REFERENCES AND NOTES

1. "Secretariat of the Convention on Biological Diversity (2020) Global Biodiversity Outlook 5 – Summary for Policy Makers" (UN Convention on Biological Diversity, 2020).
2. "Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services" S. Díaz *et al*., Eds. (IPBES secretariat, 2019).
3. D. A. Kroodsma *et al*., Tracking the global footprint of fisheries. *Science* **359**, 904–908 (2018). doi: 10.1126/science.aao5646
4. FAO, *The State of World Fisheries and Aquaculture 2020.* (FAO, 2020); www.fao.org/documents/card/en/c/ca9229en.
5. G. M. Mace *et al*., Aiming higher to bend the curve of biodiversity loss. *Nat. Sustain.* **1**, 448–451 (2018). doi: 10.1038/s41893-018-0130-0
6. T. H. Sparks *et al*., Linked indicator sets for addressing biodiversity loss. *Oryx* **45**, 411–419 (2011). doi: 10.1017/S003060531100024X
7. E. L. Hazen *et al*., Marine top predators as climate and ecosystem sentinels. *Front. Ecol. Environ.* **17**, 565–574 (2019). doi: 10.1002/fee.2125
8. B. B. Collette *et al*., High value and long life - Double jeopardy for tunas and billfishes. *Science* **333**, 291–292 (2011). doi: 10.1126/science.1208730
9. N. Pacoureau *et al*., Half a century of global decline in oceanic sharks and rays. *Nature* **589**, 567–571 (2021). doi: 10.1038/s41586-020-03173-9; pmid: 33505035
10. S. H. M. Butchart *et al*., Improvements to the Red List Index. *PLOS ONE* **2**, e140 (2007). doi: 10.1371/journal.pone.0000140; pmid: 17206275
11. IUCN Red List Categories and Criteria: Version 3.1 (IUCN Species Survival Commission, 2012); https://portals.iucn.org/library/sites/library/files/documents/RL-2001-001-2nd.pdf.
12. R. B. Sherley *et al*., Estimating IUCN Red List population reduction: JARA—A decision-support tool applied to pelagic sharks. *Conserv. Lett.* **13**, e12688 (2020). doi: 10.1111/conl.12688
13. N. K. Dulvy, S. Jennings, S. I. Rogers, D. L. Maxwell, Threat and decline in fishes: An indicator of marine biodiversity. *Can. J. Fish. Aquat. Sci.* **63**, 1267–1275 (2006). doi: 10.1139/f06-035
14. Materials and methods are available as supplementary materials.
15. Guidelines for Using the IUCN Red List Categories and Criteria, Version 14 (IUCN Standards and Petitions Committee, 2019); https://www.iucnredlist.org/resources/redlistguidelines.
16. N. K. Dulvy, H. K. Kindsvater, in *Conservation for the Anthropocene Ocean*, L. P. S, M. R. Poe, Eds. (Academic Press, 2017), pp. 339–340.
17. P. G. Fernandes *et al*., Coherent assessments of Europe's marine fishes show regional divergence and megafauna loss. *Nat. Ecol. Evol.* **1**, 0170 (2017). doi: 10.1038/s41559-017-0170
18. S. Millar, M. Dickey-Collas, Report on IUCN assessments and fisheries management approaches. ICES CM 2018/ACOM:60 (2018).
19. D. J. McCauley *et al*., Marine defaunation: Animal loss in the global ocean. *Science* **347**, 1255641 (2015). doi: 10.1126/science.1255641
20. The IUCN Red List of Threatened Species, (IUCN, 2015); www.iucnredlist.org.
21. T. J. Webb, B. L. Mindel, Global patterns of extinction risk in marine and non-marine systems. *Curr. Biol.* **25**, 506–511 (2015). doi: 10.1016/j.cub.2014.12.023; pmid: 25639240
22. R. Hilborn *et al*., Effective fisheries management instrumental in improving fish stock status. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 2218–2224 (2020). doi: 10.1073/pnas.1909726116; pmid: 31932439
23. M. G. Burgess *et al*., Protecting marine mammals, turtles, and birds by rebuilding global fisheries. *Science* **359**, 1255–1258 (2018). (2018). doi: 10.1126/science.aao4248
24. M. J. Juan-Jordá, H. Murua, H. Arrizabalaga, N. K. Dulvy, V. Restrepo, Report card on ecosystem-based fisheries management in tuna regional fisheries management organizations. *Fish Fish.* **19**, 321–339 (2017). doi: 10.1111/faf.12256
25. H. Booth, D. Squires, E. J. Milner-Gulland, The mitigation hierarchy for sharks: A risk-based framework for reconciling trade-offs between shark conservation and fisheries objectives. *Fish Fish.* **21**, 269–289 (2019). doi: 10.1111/faf.12429
26. M.-J. Juan-Jordá, N. Pacoureau, mjuanjorda/RLI_tunas_billfishes_sharks, Github (2022); https://github.com/mjuanjorda/RLI_tunas_billfishes_sharks.

ACKNOWLEDGMENTS

We thank all the authors of the fishery stock assessments used in our analysis, which were performed by the five tuna regional fisheries management organizations independent of this paper. We also thank A. B. Cooper, and the Earth to Ocean Research Group from Simon Fraser University, Canada for their highly constructive comments and feedback at the early stages of this study. **Funding:** M. J. J-J. received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement [628116], and a fellowship from "la Caixa" Foundation (ID 100010434). The "la Caixa" fellowship code is (LCF/BQ/PR20/11770005). N.K.D. was supported by the Natural Sciences and Engineering Research Council of Canada and the Canada Research Chairs Program. This paper is contribution 1130 from AZTI, Marine Research, Basque Research and Technology Alliance (BRTA). **Author contributions:** M. J. J-J. and N.K.D. conceptualized the analysis. M. J. J-J., H. M., H.A., and G. M. compiled and curated the time series data. M. J. J-J. and N.K.D. visualized the data and wrote the first draft. M. J. J-J. and N.P. conducted the statistical analysis with input from all authors. M. J. J-J. created figures with input from all authors. M. J. J-J., H. M., H.A., and N.K.D. acquired the funding. All authors discussed the time series data, developed the methodology, discussed analysis and results, contributed to writing the manuscript and oversaw the project. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Data and code used in our analysis will be available on GitHub. All other data needed to evaluate conclusions in the paper are present in the paper or the supplementary materials. **License information:** Copyright © 2022 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.sciencemag.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abj0211
Materials and Methods
Figs. S1 to S17
Tables S1 to S6
References (*27–47*)
MDAR Reproducibility Checklist

Submitted 27 April 2021; resubmitted 10 December 2021
Accepted 27 September 2022
10.1126/science.abj0211

RESEARCH ARTICLE SUMMARY

IMMUNOLOGY

Enhanced T cell effector activity by targeting the Mediator kinase module

Katherine A. Freitas, Julia A. Belk, Elena Sotillo, Patrick J. Quinn, Maria C. Ramello, Meena Malipatlolla, Bence Daniel, Katalin Sandor, Dorota Klysz, Jeremy Bjelajac, Peng Xu, Kylie A. Burdsall, Victor Tieu, Vandon T. Duong, Micah G. Donovan, Evan W. Weber, Howard Y. Chang, Robbie G. Majzner, Joaquin M. Espinosa, Ansuman T. Satpathy, Crystal L. Mackall*

INTRODUCTION: T cell immunotherapies demonstrate impressive activity against some cancers, but durable responses are not achieved in most patients. A central barrier to progress is inadequate T cell potency to eradicate large tumor burdens, which is the result of multiple factors, including T cell exhaustion, senescence, anergy, and immunosuppression. Gene editing holds promise for improving the effectiveness of cancer immunotherapy, but it remains unclear which genes, or groups of genes, will most effectively enhance T cell potency after editing. In this study, we used genome-wide CRISPR knockout screens in human T cells to identify regulators of T cell fitness.

RATIONALE: We performed two CRISPR screens in human chimeric antigen receptor (CAR) T cells using a model system that induces T cell dysfunction by mimicking chronic antigen exposure. On the basis of the hypothesis that higher rates of proliferation and cytokine production characterize the most potent antitumor T cells, we identified guide RNAs enriched in T cells that proliferate and produce both interleukin-2 (IL-2) and tumor necrosis factor–α (TNFα) after tumor exposure.

RESULTS: Both CRISPR screens identified genes that encode subunits of the Mediator complex and are contained within the Mediator kinase module. The Mediator complex acts as a bridge between enhancer-bound transcription factors and the general transcription machinery and plays a central role in establishing cellular identity by coordinating transcriptional networks. Targeted deletion of *MED12* (Mediator complex subunit 12) or *CCNC* (cyclin C) in human CAR T cells resulted in increased proliferation, cytokine production, and increased tumor clearance. Similar effects were observed with CARs targeting multiple tumor antigens and using either CD28 or 4-1BB costimulation, and in T cells expressing an engineered T cell receptor (TCR). T cells with phenotypic and transcriptomic hallmark features of stemness have demonstrated increased antitumor potency in many model systems, however, *MED12*-deficient T cells showed diminished stemness and enhanced phenotypic and transcriptomic features of effector cells. Consistent with an effector phenotype, *MED12*-deficient cells show enhanced metabolic activity and fitness, characterized by increased glycolysis, oxidative phosphorylation, and spare respiratory capacity. *MED12*-deficient T cells demonstrated sustained potency after long-term culture and repeated exposure to tumors in vitro and in vivo. Small molecule–mediated inhibition of cyclin-dependent kinases 8 and 19 (CDK8/19), the catalytic subunit of the Mediator kinase module, similarly increased expansion of healthy, nonengineered T cells.

To provide a basis for understanding these observations, we assessed changes in chromatin accessibility and modification in *MED12*-deficient T cells. Using chromatin immunoprecipitation sequencing, we demonstrated that the kinase module and core Mediator are largely colocalized in wild-type CAR T cells but loss of *MED12* increased core Mediator chromatin occupancy in more than 800 genomic regions. This is consistent with a known role for the kinase module in regulating interaction between core Mediator and RNA polymerase II (RNAPII) and led to the hypothesis that loss of *MED12* or *CCNC* in T cells selectively reduces steric hindrance between core Mediator and RNAPII, thereby increasing transcription and modulating T cell function. Consistent with this, regions with increased MED1 chromatin occupancy in *MED12*-deficient CAR T cells manifested increased H3K27 acetylation and were enriched for enhancers used by transcription factors that play a critical role in T cell fate, including several STAT (signal transducer and activator of transcription) and AP-1 (activator protein 1) family members. The most notable enhancement was observed for STAT5, which manifested as increased sensitivity to IL-2 in *MED12*-deficient T cells. Increased IL-2 sensitivity in nonengineered T cells could also be endowed by exposure to small-molecule CDK8/19 inhibitors.

CONCLUSION: These data link Mediator-induced transcriptional coactivation with T cell differentiation, identify the Mediator kinase module as a primary regulator of T cell effector programming, and demonstrate enhanced potency of *MED12*-deficient T cells in mediating antitumor effects. Technologies to inactivate genes ex vivo during cell manufacturing and in vivo are increasingly accessible, highlighting the potential for clinical translation of these findings. ■

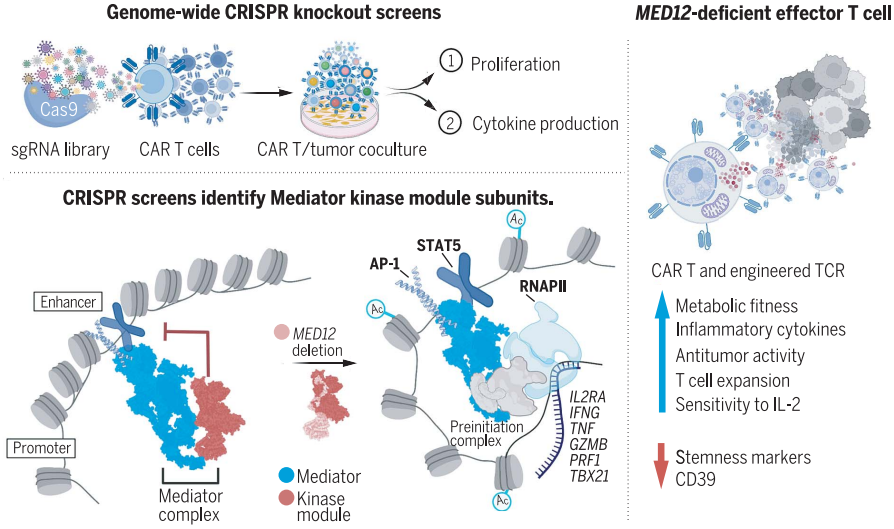


Fig. 1. Disruption of the Mediator kinase module transcriptionally rewires effector programming in human T cells. Genome-wide CRISPR screens in CAR T cells identified genes regulating T cell effector functions (top left). Targeted disruption of the Mediator kinase module increases core Mediator chromatin occupancy at enhancers used by AP-1 and STAT family transcription factors (bottom left) leading to enhanced effector function and antitumor activity (right). sgRNA, single guide RNA; Ac, acetyl.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: cmackall@stanford.edu
Cite this article as K. A. Freitas *et al.*, *Science* **378**, eabn5647 (2022). DOI: 10.1126/science.abn5647

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abn5647

RESEARCH ARTICLE

IMMUNOLOGY

Enhanced T cell effector activity by targeting the Mediator kinase module

Katherine A. Freitas[1,2]†, Julia A. Belk[3]†, Elena Sotillo[2], Patrick J. Quinn[2], Maria C. Ramello[2], Meena Malipatlolla[2], Bence Daniel[4,5], Katalin Sandor[5], Dorota Klysz[2], Jeremy Bjelajac[2,6], Peng Xu[2], Kylie A. Burdsall[2], Victor Tieu[7], Vandon T. Duong[7], Micah G. Donovan[8], Evan W. Weber[2,9]‡, Howard Y. Chang[9,4,10], Robbie G. Majzner[2,11], Joaquin M. Espinosa[8,12], Ansuman T. Satpathy[2,9,5]§, Crystal L. Mackall[2,9,11,13]§*

T cells are the major arm of the immune system responsible for controlling and regressing cancers. To identify genes limiting T cell function, we conducted genome-wide CRISPR knockout screens in human chimeric antigen receptor (CAR) T cells. Top hits were *MED12* and *CCNC*, components of the Mediator kinase module. Targeted *MED12* deletion enhanced antitumor activity and sustained the effector phenotype in CAR- and T cell receptor–engineered T cells, and inhibition of CDK8/19 kinase activity increased expansion of nonengineered T cells. *MED12*-deficient T cells manifested increased core Meditator chromatin occupancy at transcriptionally active enhancers—most notably for STAT and AP-1 transcription factors—and increased *IL2RA* expression and interleukin-2 sensitivity. These results implicate Mediator in T cell effector programming and identify the kinase module as a target for enhancing potency of antitumor T cell responses.

T cell–based immunotherapies, including immune checkpoint inhibitors and adoptive cell therapies, have demonstrated impressive antitumor effects in many cancers (*1*–*8*), but durable responses are not achieved in most patients. A central barrier to progress is limited T cell potency, resulting from a myriad of factors, including T cell exhaustion, senescence, anergy, and local and systemic immunosuppression (*9*–*12*). Advances in understanding the biology of T cell exhaustion are providing novel approaches to prevent these phenomena, including overexpression of c-JUN (*13*), deletion of nuclear receptor subfamily 4A (NR4A) (*14*), and transient induction of T cell rest (*15*, *16*). However, it remains unclear whether exhaustion resistance will be sufficient to overcome the multitude of immunosuppressive factors in the tumor microenvironment.

Gene editing technologies are providing unparalleled opportunities to engineer more potent human T cells. Ex vivo CRISPR has been used safely to deliver gene-edited tumor-specific T cells to humans with cancer (*17*), and the CRISPR platform has been optimized to conduct forward genetic screens in primary human T cells to identify novel targets to augment T cell function. These approaches have identified genes that regulate programmed cell death protein 1 (PD-1) expression, T cell proliferation and persistence, and resistance to adenosine-mediated immunosuppression (*18*–*20*), but work is ongoing to define genes for which editing will most potently augment antitumor responses in humans. In this study, we used CRISPR screening to identify genes that regulate effector function in primary human T cells expressing chimeric antigen receptors (CARs) and discovered that *MED12* (Mediator complex subunit 12) and *CCNC* (cyclin C), genes encoding proteins in the kinase module of the Mediator complex, negatively regulate T cell effector activity. Mediator, an evolutionarily conserved multisubunit protein complex that acts as a bridge between enhancer-bound transcription factors and the general transcription machinery, is required for gene transcription and plays a central role in establishing cellular identity by coordinating transcriptional networks (*21*–*23*). Across multiple CAR T cell models with different costimulatory domains, and in T cells expressing an engineered T cell receptor (TCR), we discovered that genetic disruption of the kinase module of Mediator induced transcriptional and epigenetic changes that resulted in enhanced effector function, metabolic fitness, and increased antitumor activity. Small molecule–mediated inhibition of cyclin-dependent kinases 8 and 19 (CDK8/19) in nonengineered T cells phenocopied several of these enhancements. These results implicate the Mediator kinase module as a therapeutic target for augmenting T cell fitness and identify a previously unknown role for *MED12* in regulating human T cell function.

Results

Genome-wide screen identifies the Mediator kinase module as a regulator of CAR T cell expansion and cytokine production

To identify genes that restrain CAR T cell function, we performed two genome-wide CRISPR deletion screens to identify negative regulators of T cell expansion and cytokine production in primary T cells from two donors transduced with HA-28ζ, a high-affinity GD2-targeting CAR that induces functional, transcriptomic, and epigenetic hallmarks of T cell exhaustion (*13*, *16*). Using a previously published single guide RNA (sgRNA) library (*24*), editing was achieved by adapting the SLICE (sgRNA lentiviral infection with Cas9 protein electroporation) platform (*18*) to incorporate CAR transduction (Fig. 1A and fig. S1A). We detected 98% of the sgRNA library in transduced CAR T cells, with 19,885 genes targeted by at least four sgRNAs (fig. S1, B and C). Successful editing was confirmed by drop out of sgRNAs targeting a "gold standard" set of essential genes but not control guides after 23 days in culture (fig. S1D and tables S1 and S2).

For the expansion screen, we cultured the transduced cells in vitro for 15 days, then co-cultured with $GD2^+$ tumor cells until day 23 and compared sgRNA abundance between day 0 and day 23 (fig. S1E). Per the MAGeCK algorithm (*25*), both donors showed enrichment of sgRNAs targeting genes known to inhibit T cell survival, such as *FAS* and *CASP3* (*26*) (Fig. 1B), whereas sgRNAs targeting genes known to promote T cell proliferation, such as *IL2RG*, *MYC*, and *ZAP70*, were depleted. The top hits in the expansion screen were *CCNC* (cyclin C) and *MED12* (Mediator complex subunit 12), members of the kinase module of Mediator, with seven of seven guides targeting *CCNC* and eight of eight guides targeting *MED12* positively enriched (fig. S2, A to C). The expansion screen also showed depletion of sgRNAs targeting *BATF* and *JUNB*, suggesting a survival role for activator protein 1 (AP-1) family members in the setting of chronic stimulation.

In the cytokine production screen, we cultured HA-28ζ CAR T cell knockout libraries in vitro

[1]Immunology Graduate Program, Stanford University School of Medicine, Stanford, CA, USA. [2]Center for Cancer Cell Therapy, Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA. [3]Department of Computer Science, Stanford University, Stanford, CA, USA. [4]Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA, USA. [5]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA. [6]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, CA, USA. [7]Department of Bioengineering, Stanford University School of Medicine, Stanford, CA, USA. [8]Department of Pharmacology, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [9]Parker Institute for Cancer Immunotherapy, San Francisco, CA, USA. [10]Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA. [11]Division of Pediatric Hematology/Oncology and Division of Stem Cell Transplantation and Regenerative Medicine, Department of Pediatrics, Stanford University School of Medicine, Stanford, CA, USA. [12]Linda Crnic Institute for Down Syndrome, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [13]Division of Blood and Marrow Transplantation and Cell Therapy, Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA.
***Corresponding author. Email: cmackall@stanford.edu**
‡Present address: Department of Pediatrics, University of Pennsylvania, Philadelphia, PA, USA.
†These authors contributed equally to this work.
§These authors contributed equally to this work.

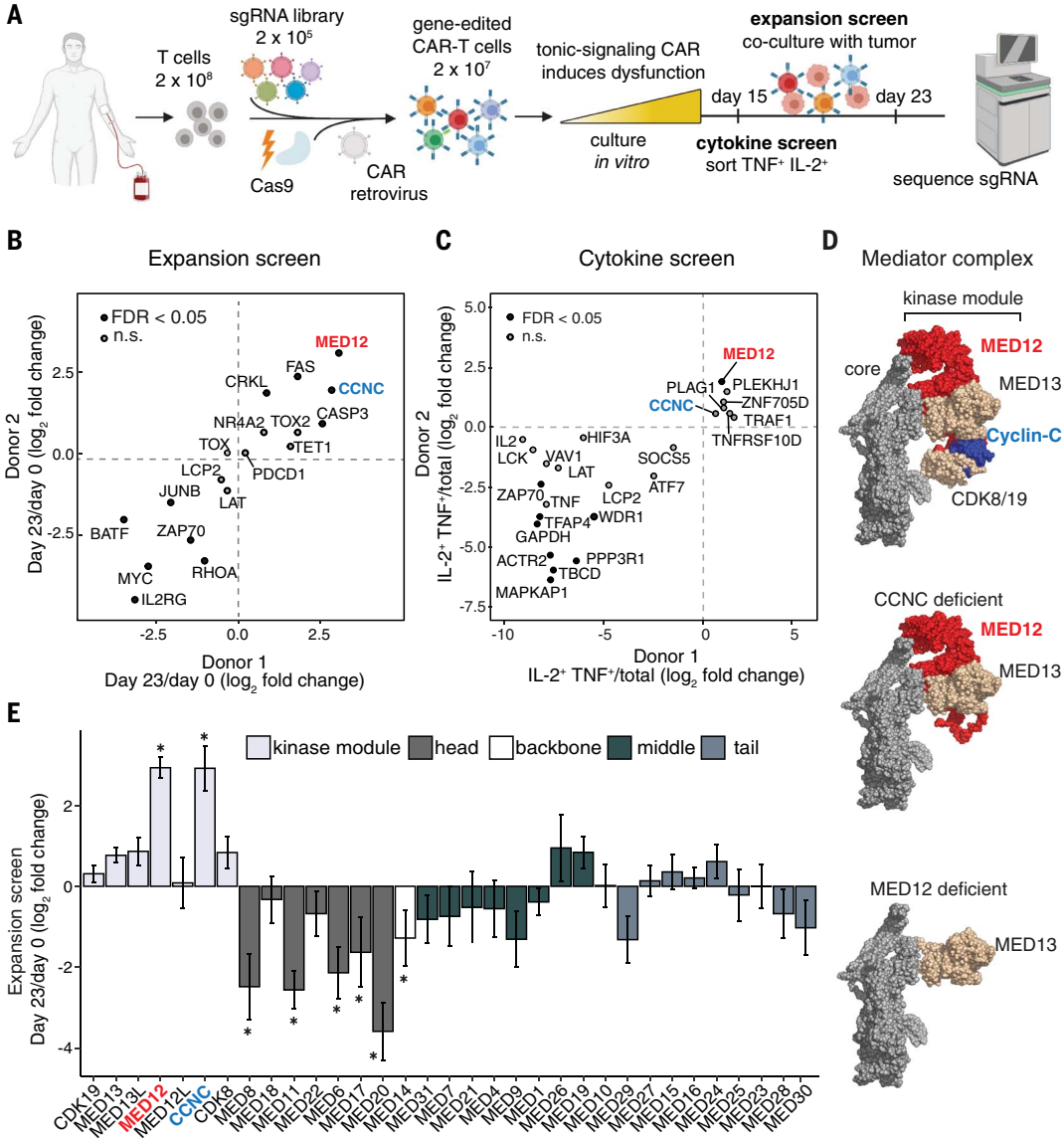


Fig. 1. Genome-wide CRISPR screen identifies subunits of the Mediator kinase module as regulators of CAR T cell effector function. (**A**) Schematic depicting CRISPR knockout screen for regulators of cytokine production and CAR T cell expansion using a tonic signaling model of CAR T cell exhaustion. (**B**) Enrichment of gene knockouts in replicate expansion screens. CRISPR-edited HA-28ζ CAR T cells were generated from two donors, cultured in vitro for 15 days, and then cocultured with GD2$^+$ tumor cells until day 23. n.s., not significant. (**C**) Enrichment of gene knockouts in replicate cytokine production screens. CRISPR-edited HA-28ζ CAR T cells were generated from two donors, cultured in vitro for 15 days, stimulated with GD2$^+$ tumor cells, and the top 10% of TNFα- and IL-2–expressing cells were isolated by FACS. (**D**) Predicted cryo-EM structure of yeast Mediator complex (top) showing the effect that depletion of cyclin C ("CCNC deficient") or *MED12* ("MED12 deficient") would have on assembly of other subunits. Core Mediator is shown in gray, the kinase module is colored. Representations were created with Chimera using Protein Data Bank IDs 7KPX and 5U0P. (**E**) Bar graphs depicting enrichment of sgRNA targeting all Mediator complex subunits in the expansion screen. Data are mean ± SD (n = 4 to 10 guides per gene). Colors indicate module of the Mediator complex assigned to each subunit. [(B), (C), and (E)] Data are mean of n = 4 to 10 guides per gene. Data are pooled from two independent experiments (n = 2 donors). Gene-level statistical significance was determined by the MAGeCK algorithm. *FDR < 0.05.

for 15 days, added GD2$^+$ tumor cells to the culture for 6 hours, used fluorescence-activated cell sorting (FACS) to sort cells, and compared sgRNA abundance in CAR T cells expressing interleukin-2 (IL-2) and tumor necrosis factor–α (TNFα) protein against the total population (fig. S1E). *MED12* was enriched in the cytokine screen, with six of seven guides demonstrating enrichment, while several genes, including nine of nine guides targeting *ZAP70* were depleted, consistent with the known role for *ZAP70* in CAR signaling (*27*) (Fig. 1C and fig. S2, B and C). Six of nine guides targeting *TNF* and five of eight guides targeting *IL2* were also significantly depleted, demonstrating an expected loss of cytokine-negative cells in the sorted population (fig. S2B).

Mediator consists of a 26-subunit core organized into head, middle, backbone, and tail domains, and a four-subunit dissociable kinase module (*28*, *29*) (Fig. 1D). *CCNC* and *MED12* are both centrally located in the kinase module (*30*), suggesting that loss of either gene disrupts a common function. Consistent with this, sgRNAs targeting all members of the kinase module were positively enriched in the expansion screen except for *MED12L*, which is not expressed in T cells (Fig. 1E and fig. S2D). In contrast, sgRNAs targeting subunits of the head, backbone, and middle domains of core Mediator were associated with poor expansion (Fig. 1E). Together, these data demonstrate a requirement for the core Mediator complex in T cell survival and a regulatory role for the Mediator kinase module in T cell expansion and cytokine production.

Mediator kinase module–deficient CAR T cells demonstrate increased in vitro and in vivo expansion independent of costimulation domain or tonic CAR signaling

To validate the expansion screen findings, sgRNAs targeting *CCNC*, *MED12*, or *AAVS1* as a control were delivered as ribonucleoprotein 3 days after T cell activation followed by retroviral transduction of the HA-28ζ CAR (fig. S3A). *CCNC* and *MED12* deletion were confirmed by immunoblotting and Sanger sequencing using Inference of CRISPR Edits (ICE) (*31*) (fig. S3, B and C). Because CAR transduction efficiency, as well as the ratio of CD4$^+$ to CD8$^+$

cells, could affect CAR T cell function, we confirmed that loss of *MED12* or *CCNC* did not change CAR expression or the ratio of $CD4^+$ to $CD8^+$ cells, nor did it affect retroviral integration of green fluorescent protein (GFP) (fig. S3, D to H). *MED12*- and *CCNC*-deficient HA-28ζ CAR T cells showed greater expansion than did control cells over 23 days in culture (Fig. 2A). We previously reported that HA-28ζ CAR T cells develop hallmark features of exhaustion attributable to tonic signaling (*13*, *16*, *32*). To determine whether MED12 and/or CCNC would limit functionality of CAR T cells that do not develop exhaustion in vitro, we deleted *MED12* and *CCNC* in T cells expressing the CD19-28ζ CAR (*13*, *16*, *32*). *CCNC*- and *MED12*-deficient CD19-28ζ CAR T cells demonstrated increased expansion compared with control cells over 23 days in culture and after serial stimulation with tumor cells (Fig. 2A and fig. S4A). Furthermore, adoptively transferred *MED12*- and *CCNC*-deficient HA-28ζ and CD19-28ζ CAR T cells showed increased in vivo expansion in tumor-bearing NSG mice compared with control CAR T cells (fig. S4B). We also observed enhanced expansion of *MED12*- and *CCNC*-deficient HER2-4-1BBζ CAR T cells (fig. S4B), confirming that the findings are not restricted to CAR T cells incorporating a CD28 costimulatory domain. Together, these results demonstrate that *MED12*- and *CCNC*-deficient human CAR T cells manifest enhanced antigen-driven expansion regardless of whether the CAR incorporates a CD28 or 4-1BB costimulatory domain or whether the CAR T cells manifest hallmark features of exhaustion.

To determine whether these effects are dependent on catalytic activity of the kinase module, we cultured healthy human T cells following anti-CD3/CD28 activation with compounds that are dual inhibitors of CDK8 and CDK19 and observed significant increases in T cell expansion (Fig. 2B). Additionally, overexpression of *MED12* suppressed T cell proliferation, confirming that the kinase module restrains T cell expansion (fig. S4, C and D). *MED12* and *CCNC* behave as tumor suppressors in some settings (*33*, *34*), however, when we removed IL-2 from the culture medium, we observed a complete loss of viable cells within 3 weeks (fig. S4E), indicating that expansion is not associated with transformation, as the cells remain IL-2 dependent. Together, these results demonstrate that the kinase module is a potent modulator of human T cell expansion.

Mediator kinase module–deficient CAR T cells produce higher levels of inflammatory cytokines after antigen stimulation

Next, we sought to confirm the effects of *MED12* and *CCNC* deletion on antigen-induced cytokine production using bulk assays and by assessing single-cell production of IL-2 and TNFα by flow cytometry. Bulk cultures of antigen-stimulated *MED12*- and *CCNC*-deficient HA-28ζ, CD19-28ζ, and HER2-4-1BBζ CAR T cells produced higher levels of cytokines and showed increased frequencies of IL-2 and TNFα–expressing cells (Fig. 2C and fig. S5, A and B). Additionally, we found elevated *IL2*, *IFNG*, and *TNF* mRNA levels in *MED12*-deficient cells compared with control cells, indicating that these changes are transcriptionally mediated (fig. S5C). To assess the impact of Mediator kinase module disruption more broadly on antigen-induced cytokine secretion, we performed bead-based multiplex immunoassay profiling of 38 cytokines in supernatants collected from CD19-28ζ CAR T cells stimulated with $CD19^+$ Nalm6 leukemia cells for 24 hours. Hierarchical clustering showed that the cytokine profile of *MED12*- and *CCNC*-deficient CD19-28ζ CAR T cells was distinct from that of controls (Fig. 2D), with increased proinflammatory cytokines including interferon-γ (IFNγ), TNFα, IL-17, and IL-6; increased inflammatory chemokines CXCL10 and CCL3; and increased common gamma chain family cytokines IL-2 and IL-9, which promote T cell survival and differentiation (fig. S5D). Together, these results demonstrate that *MED12* and *CCNC* constrain antigen-induced T cell expansion and inflammatory cytokine production and raise the prospect that *MED12*- or *CCNC*-deficient T cells may demonstrate enhanced antitumor immune responses.

Mediator kinase module–deficient CAR T cells demonstrate increased metabolic fitness and antitumor activity

To assess whether *MED12*- or *CCNC*-deficient T cells manifest metabolic features of enhanced effector functionality (*35*), we measured glycolytic and oxygen consumption rates. We observed increased basal and maximal oxygen consumption in *MED12*- and *CCNC*-deficient cells, despite no change in mitochondrial mass, and increased basal and maximal rates of glycolysis (Fig. 2, E and F, and fig. S6, A and B). Stimulation of CD19-28ζ CAR T cells via the CAR resulted in higher levels of pS6 in *MED12*-deficient cells, demonstrating enhanced antigen-dependent activation of the mammalian target of rapamycin complex 1(MTORC1) pathway which may contribute to the increased metabolic activity observed (fig. S6, C and D). The simultaneous increases in rates of oxidative phosphorylation and glycolysis observed in *MED12*-deficient cells is similar to the metabolic state described in early activated T cells, which have not yet differentiated into short-lived effector cells or memory cells (*36*).

CCNC- and *MED12*-deficient CAR T cells demonstrated enhanced antitumor function in vivo, in a model wherein NSG mice were inoculated with Nalm6 or Nalm6-GD2 leukemic cells and treated 3 days later with gene-edited CD19-28ζ or HA-28ζ CAR T cells, respectively (Fig. 2, G and H, and fig. S6, E and F). Similarly, we observed enhanced tumor control and prolonged survival in mice engrafted with 143B osteosarcoma cells and treated with *MED12*- and *CCNC*-deficient CAR T cells expressing the HER2-4-1BBζ receptor (Fig. 2I and fig. S6G). Collectively, these results demonstrate that *MED12*- and *CCNC*-deficient T cells manifest enhanced hallmark features of effector cells, spanning antigen-induced expansion, cytokine production, metabolic fitness, and killing capacity.

MED12-deficient T cells demonstrate sustained effector function under chronic stimulation

To further investigate the impact of kinase module disruption on longer-term T cell fitness and following repetitive antigen stimulation, we focused on *MED12* because this gene was a top hit in both CRISPR screens. *MED12*-deficient HA-28ζ CAR T cells demonstrated increased expansion and cytotoxicity after repeated stimulation with Nalm6-GD2 tumor cells (Fig. 3, A and B, and fig. S7, A and B). Similarly, *MED12*-deficient HA-28ζ CAR T cells cultured in vitro until day 54 continued to demonstrate enhanced T cell expansion and cytokine production (Fig. 3, C and D). To further characterize the *MED12*-deficient phenotype at late time points, we performed single-cell proteomic analysis of 34 proteins using mass cytometry to measure lineage-defining transcription factors and cell surface markers associated with activation, exhaustion, and T cell differentiation (table S3). Both control and *MED12*-deficient HA-28ζ CAR T cells displayed an activated effector phenotype after 50 days in culture, with high expression of Ki67, CD69, Tbet, TOX, CD25, and CD122 and low expression of IL7R, CD45RA, CD27, CD28, CCR7, and TCF7 relative to non-activated T cells isolated from peripheral blood (fig. S7, C and D). Relative to control HA-28ζ cells, loss of *MED12* substantially reduced expression of CD39, a marker associated with terminal exhaustion and diminished stemness (Fig. 3, E and F) (*37*, *38*), while PD-1 and TIM3 were unchanged and LAG3 was elevated (fig. S7E). Additionally, *MED12*-deficient cells maintained an elevated $CD8^+/CD4^+$ ratio during long-term culturing (fig. S7F).

To determine whether cells lacking *MED12* would manifest sustained antitumor activity in vivo, we engrafted mice with Nalm6-GD2 leukemia, treated them with *MED12*-deficient or control HA-28ζ CAR T cells, and rechallenged them with additional tumor cells 26 days after CAR T cells were administered. *MED12*-deficient CAR T cells protected from rechallenge with $GD2^+$ but not $GD2^-$ tumor cells, demonstrating prolonged antigen-specific antitumor activity (Fig. 3, G to I). Together, these results show that loss of MED12 results in long-term

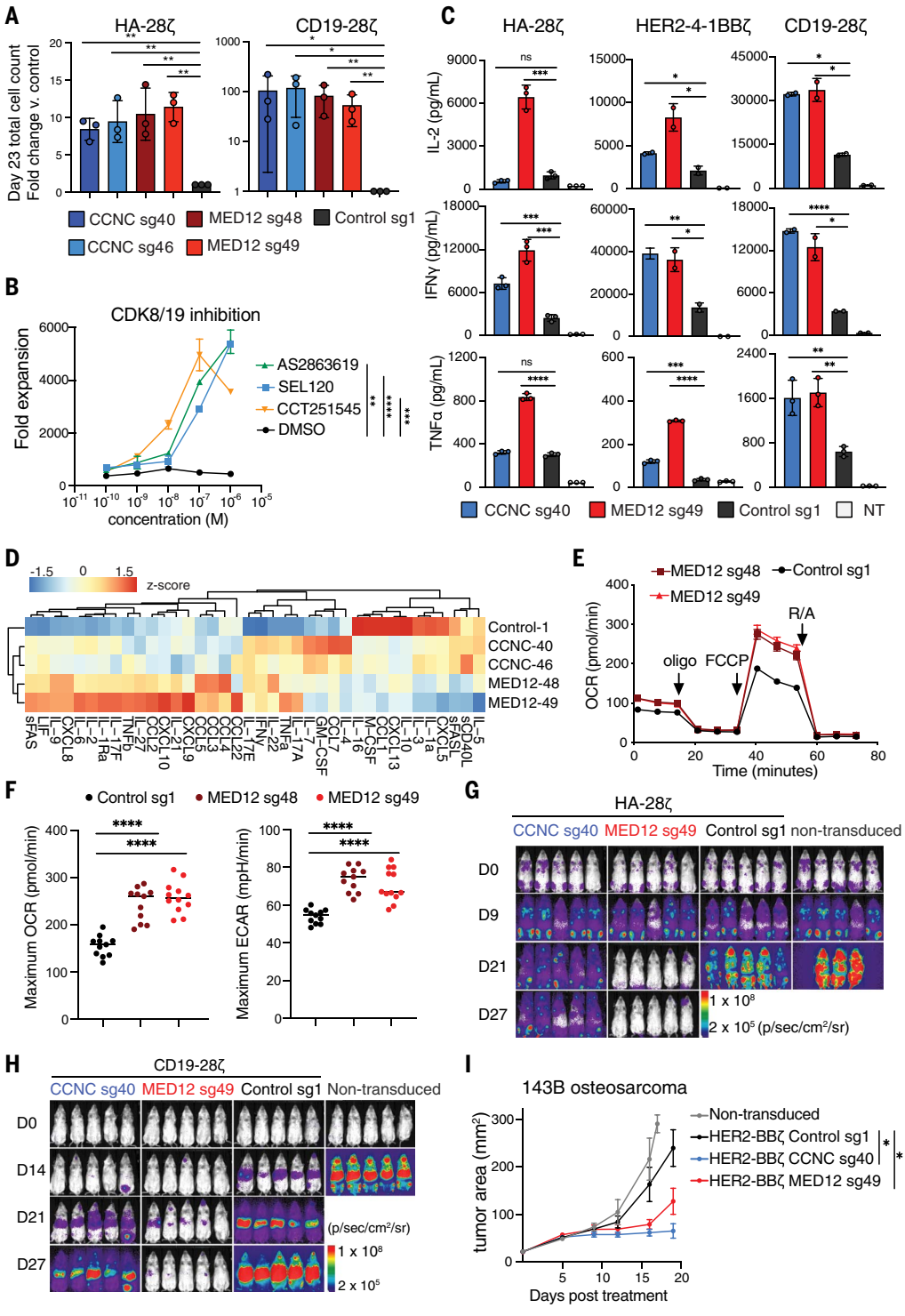


Fig. 2. Disruption of the Mediator kinase module in CAR T cells enhances T cell effector function and tumor clearance. (**A**) In vitro T cell expansion of *CCNC*- and *MED12*-deficient HA-28ζ (left) and CD19-28ζ (right) CAR T cells. Fold change in total cell count after 23 days in culture relative to control CAR T cells edited at the safe harbor AAVS1 locus. Two different sgRNAs were used to validate each candidate gene. Data are mean ± SD of $n = 3$ donors. Ratio paired t test. $^*P < 0.05$, $^{**}P < 0.01$. (**B**) In vitro expansion of human primary T cells with dual inhibitors of CDK8 and CDK19 over 15 days in culture. Inhibitors were supplemented to the media every 48 hours. Data are mean ± SD of $n = 2$ replicate wells. Representative of three independent experiments. (**C**) IL-2 (top), IFNγ (middle), and TNFα (bottom) cytokine release after 24-hour coculture with tumor cells from non-transduced (NT) and CAR T cells edited with sgRNAs targeting *MED12*, *CCNC*, or safe-harbor control. HA-28ζ, CD19-28ζ, and HER2-4-1BBζ CAR T cells were stimulated 1:1 with Nalm6-GD2, Nalm6, or 143B cells, respectively. Data are mean ± SD from duplicate or triplicate wells. Representative results of $n = 4$ donors (HA-28ζ and CD19-28ζ) or $n = 2$ donors (HER2-4-1BBζ). Non-transduced T cells were activated with CD3/28 stimulation but were not transduced with retrovirus or gene edited. (**D**) Heatmap of 38 cytokines produced by control, *CCNC*-deficient, or *MED12*-deficient CD19-28ζ CAR T cells after 24-hour coculture with Nalm6 leukemia cells. Data are mean from duplicate wells in a multiplex bead-based assay. Two different sgRNAs were used to validate each candidate gene. (**E** and **F**) Metabolic rate as measured by Seahorse analysis of oxygen consumption rate (OCR) and extracellular acidification rate (ECAR) of control or *MED12*-deficient CD19-28ζ CAR T cells under resting and challenge conditions. Data are mean of $n = 12$ replicate wells. Representative results from two independent experiments. (**G**) Analysis of tumor clearance. NSG mice were injected intravenously with 1.0×10^6 Nalm6-GD2 leukemia cells and treated with 2.0×10^5 nontransduced or *CCNC*- or *MED12*-deficient HA-28ζ CAR T cells 9 days after tumor infusion ($n = 5$ mice). (**H**) Analysis of tumor clearance. NSG mice were injected intravenously with 1.0×10^6 Nalm6 leukemia and treated with 2.5×10^5 nontransduced or *CCNC*- or *MED12*-deficient CD19-28ζ CAR T cells 3 days after tumor infusion ($n = 5$ mice). (**I**) Analysis of tumor clearance. Tumor area of NSG mice injected intramuscularly with 1×10^6 143B osteosarcoma cells and treated 4 days later with 5×10^6 nontransduced or *CCNC*- or *MED12*-deficient HER2-4-1BBζ CAR T cells. Data are mean ± SD of $n = 5$ mice (nontransduced, *MED12*-deficient, and *CCNC*-deficient) or $n = 4$ mice (control). Two-way analysis of variance (ANOVA) test with Dunnett's multiple comparison test. $^*P < 0.01$. [(G) to (I)] Representative experiment from two independent experiments ($n = 2$ donors). [(B), (C), and (F)] Two-tailed unpaired Student's t test. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

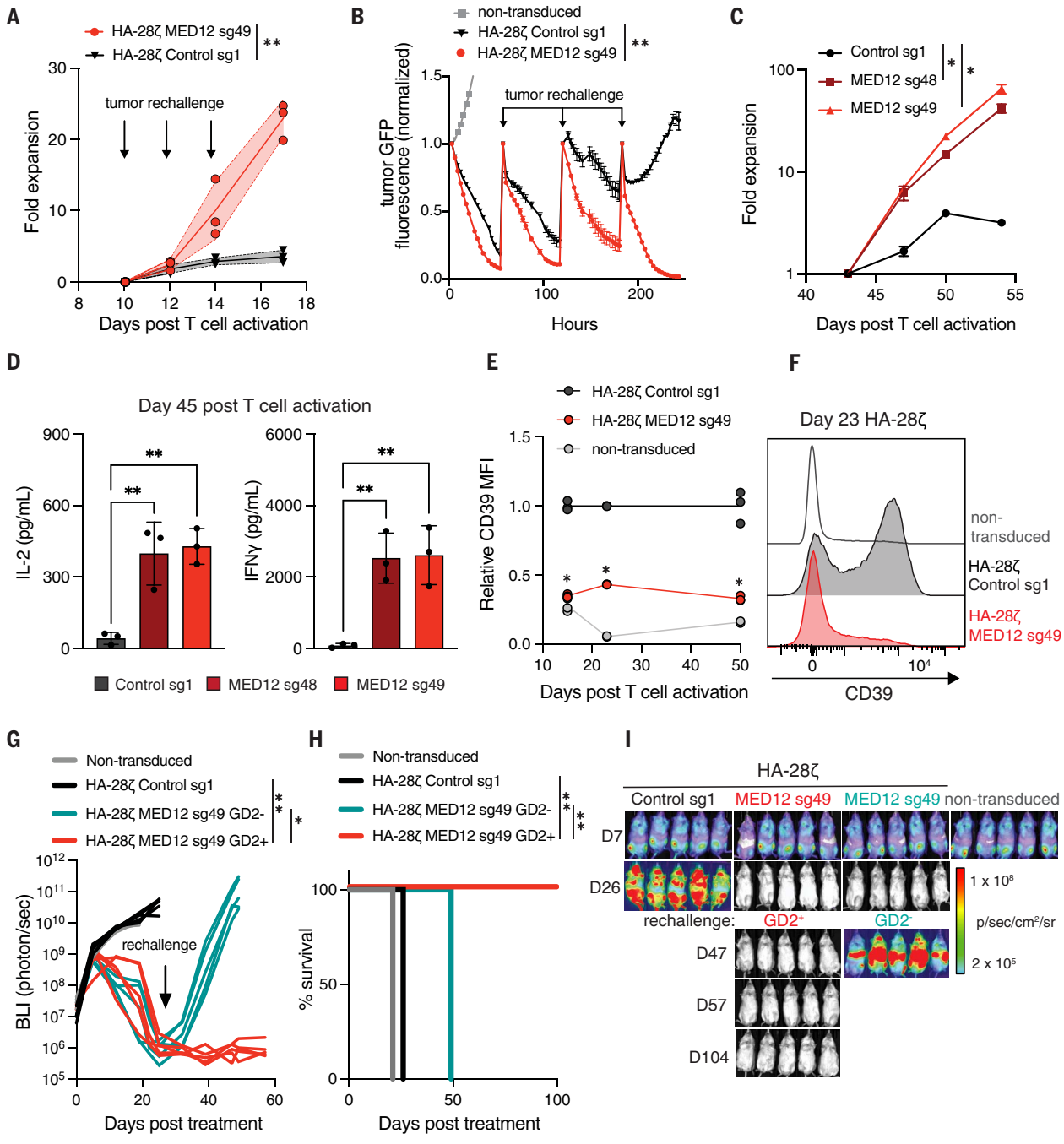


Fig. 3. Loss of *MED12* sustains effector function during chronic stimulation. (**A**) Antigen-driven in vitro expansion of control and *MED12*-deficient HA-28ζ CAR T cells. CAR T cells were serially stimulated with GD2$^+$ tumor cells in the absence of IL-2 at 48-hour intervals at a 1:1 effector to target cell ratio. Data are mean ± SD of n = 3 donors. (**B**) Cytotoxicity of control and *MED12*-deficient HA-28ζ CAR T cells against GFP$^+$ Nalm6-GD2 leukemia after serial stimulation beginning 10 days after T cell activation. Cells were counted and replated at a 1:1 ratio of T cells to tumor cells at 48- to 72-hour intervals in media without IL-2. Data are mean ± SD of n = 3 replicate cultures. Representative of three independent experiments. (**C**) In vitro expansion of control and *MED12*-deficient HA-28ζ CAR T cells cultured with IL-2. Two different sgRNAs were used to validate each candidate gene. Data are mean ± SD of n = 2 replicate cultures. (**D**) IL-2 (left) and IFNγ (right) release after 24-hour coculture with tumor cells 45 days after T cell activation. Control and *MED12*-deficient HA-28ζ CAR T cells were stimulated 1:1 with Nalm6-GD2. Data are mean ± SD from n = 3 replicate cultures. (**E** and **F**) Flow cytometric analysis of CD39 expression in non-transduced or HA-28ζ CAR T cells. Mean fluorescence intensity is normalized to the control at each time point. Data are mean ± SD of n = 3 replicate wells. Statistical comparison is between control and *MED12*-deficient CAR T cells. (**G** and **I**) Analysis of tumor clearance. Bioluminescent imaging (BLI) of NSG mice injected intravenously with 1 × 10^6 Nalm6-GD2 leukemia and treated with 4 × 10^5 nontransduced or HA-28ζ CAR T cells 7 days after tumor infusion and rechallenged 26 days later with Nalm6 or Nalm6-GD2 cells (n = 5 mice). (**H**) Survival of CAR-treated mice shown in (G). Survival curves were compared with the Log-rank Mantel-Cox test. *P < 0.01. [(A) to (D)] Two-tailed unpaired Student's t test. *P < 0.05, **P < 0.01. [(E) and (G)] Two-way ANOVA test with Dunnett's multiple comparison test. *P < 0.01. [(C) to (H)] Representative of two independent experiments.

enhancement of T cell fitness, both in the setting of chronic stimulation due to the tonic signaling HA-28ζ CAR and after repeated encounters with tumor cells.

Loss of MED12 increases effector function in T cells using a TCR for tumor recognition

To determine whether the effects observed in *MED12*-deficient CAR T cells were generalizable to T cells that use a TCR for target recognition, we deleted *MED12* and transduced cells with the α and β chains of a TCR that recognizes New York esophageal squamous cell carcinoma (NY-ESO-1), a tumor antigen found in numerous human cancers, including melanoma and synovial sarcoma (*39*). Compared with controls, *MED12*-deficient NY-ESO-1 T cells demonstrated an increased proportion of cells bearing an effector memory phenotype, and a lower proportion bearing a stem cell memory phenotype (Fig. 4, A and B, and fig. S8A). *MED12*-deficient NY-ESO-1 T cells showed reduced expression of CD45RA and IL7R and elevated expression of Ki67, IL2RA, ICOS, and Tbet, consistent with an activated, proliferating effector phenotype with diminished quiescence compared with control cells (Fig. 4, C to E, and fig. S8, B and C).

Loss of *MED12* resulted in increased T cell expansion in culture and increased cytokine release upon coculture with NY-ESO-1$^+$ melanoma cells (Fig. 4, F and G). To determine whether loss of *MED12* increased antitumor activity in vivo, we engrafted mice with A375 melanoma and treated them with control or *MED12*-deficient NY-ESO-1 T cells. Complete tumor clearance was observed 28 days after treatment in seven of nine mice in the *MED12*-deficient group, while zero of nine mice were tumor-free in the control group (Fig. 4, H and I, and fig. S8D). Single-cell transcriptomic profiling of tumor-infiltrating NY-ESO-1 T cells showed that *MED12*-deficient cells expressed higher levels of genes encoding cytotoxic molecules including perforin, granzyme B, and IFNγ and lower levels of natural killer (NK) cell receptors *KLRD1* and *KLRB1*, which have been associated with T cell dysfunction during chronic antigen exposure (Fig. 4, J and K, and fig. S8E) (*40*), and cell cycle analysis showed that a higher fraction of *MED12*-deficient tumor-infiltrating lymphocytes were in S phase, consistent with increased proliferation in tumors (fig. S8F).

To assess similarities between *MED12*-deficient T cells and naturally occurring T cell populations found in tumors from human donors, we compared the single-cell RNA sequencing (scRNA-seq) profiles of control and *MED12*-deficient NY-ESO-1 T cells isolated from xenograft tumors to a set of 17 previously described T cell gene expression signatures (*41*). *MED12*-deficient T cells were enriched for the interferon stimulation gene signature and depleted in the NK-like T cell signature, consistent with our previous observations. We also found modest enrichment of the GZMK$^+$ early T effector memory signature, which supports the model that loss of *MED12* sustains a transitory effector memory phenotype that precedes terminal effector differentiation (fig. S8, G and H). Together, the data demonstrate phenotypic differences and functional enhancements observed in *MED12*-deficient CAR T cells are generalizable to T cells expressing an engineered TCR and indicate that disruption of the kinase module could have broad utility in T cell–directed immunotherapies.

MED12-deficient CAR T cells have an effector-like phenotype and display an activated transcriptional program

MED12-deficient CD19-28z CAR T cells displayed effector phenotypes on the basis of an absence of CCR7; however, *MED12*-deficient CAR T cells expressed high levels of CD45RO, whereas control cells were largely negative for this marker at this time point (*42*–*44*) (Fig. 5, A and B). Both *MED12*-deficient and control CD19-28ζ CAR T cells expressed high levels of Blimp-1 and low levels of Eomes and CD28, consistent with an effector phenotype (*45*); however, unbiased clustering demonstrated significant distinctions between *MED12*-deficient and control phenotypes (Fig. 5C and fig. S9A). *MED12*-deficient cells expressed higher levels of T-bet, ICOS, TOX, and CD45RO and lower levels of CD45RA and IL7R, a phenotype previously associated with short-lived effector cells (SLECs) that have not undergone terminal differentiation (*46*) (Fig. 5D and fig. S9, B and C). Paradoxically, *MED12*-deficient cells also expressed high levels of CD62L, which is usually associated with stem cell and central memory subsets and not typically expressed by SLECs. *MED12*-deficient CD19-28ζ CAR T cells also expressed high levels of LAG-3, whereas other exhaustion markers such as PD-1, TIM3, TIGIT, and CD39 were lowly expressed and unchanged from control cells (Fig. 5D and fig. S9C). Together, these results demonstrate that *MED12*-deficient CAR T cells manifest expansion of a distinctive CCR7$^-$IL7R$^-$Tbet$^+$ICOS$^+$CD62L$^+$ effector cell subset that displays enhanced cytokine production, effector cell potency, metabolic fitness, and antitumor activity.

To assess genome-wide transcriptional differences between *MED12*-deficient and control cells, bulk RNA-seq was performed on day 15 CD19-28ζ CAR T cells after a 3-hour stimulation via the CAR in vitro. *MED12*-deficient cells were transcriptionally distinct from control cells, with differential expression of 3501 genes between genotypes in at least one condition (Fig. 5E). Consistent with the functional and phenotypic data, *MED12*-deficient cells demonstrated increased expression of numerous genes associated with effector cell differentiation, including AP-1 family transcription factors (*FOS*, *JUNB*, *BATF*, *BATF3*), *IFNG*, *TNF*, *CD38*, *IL2RA*, and *CD69*. They also expressed lower levels of genes associated with T cell stemness, including *LEF1*, *TCF7*, *CD27*, and *IL7R*, and decreased expression of genes associated with T cell quiescence, including *KLF2* and *FOXO1*. Consistent with *MED12*-deficient cells manifesting enhanced cytokine secretion and metabolic fitness, gene set enrichment analysis (GSEA) of differentially expressed genes revealed enrichment of metabolic and cytokine-related gene sets (Fig. 5, F and G). Transcriptional changes induced by loss of MED12 were largely shared between CD4$^+$ and CD8$^+$ T cell subsets (fig. S9, D to F). Together, these results indicate that loss of MED12 promotes a transcriptional program consistent with an activated effector memory phenotype with enhanced metabolic fitness and cytokine secretion capacity.

Loss of MED12 increases core Mediator chromatin occupancy at transcriptionally active enhancers

The Mediator complex lacks a DNA binding domain but interacts with chromatin through protein-protein interactions with DNA-bound transcription factors and RNA polymerase II (RNAPII) (*47*). To identify the genomic locations of chromatin–Mediator interactions, we performed chromatin immunoprecipitation sequencing (ChIP-seq) using antibodies against MED12 and MED1 to profile chromatin binding of the kinase module and core Mediator, respectively, in the presence or absence of MED12. Comparison of MED1- and MED12-bound genomic regions in control CD19-28ζ CAR T cells showed MED1 and MED12 colocalization in 86.1% of sites, whereas MED1 was found exclusively at 13.1% of sites and MED12 was found exclusively at only 0.7% of sites (Fig. 6A). These results demonstrate that, in human T cells, the kinase module rarely contacts chromatin in the absence of core Mediator and the kinase module is present at the majority of sites occupied by core Mediator. Given the degree of colocalization observed, we predicted that *MED12* deletion could have widespread effects on transcription and function of the core Mediator.

To assess the effect of *MED12* deletion on core Mediator chromatin occupancy, we compared genomic regions bound by MED1 in control and *MED12*-deficient CAR T cells. Principal components analysis showed global differences in MED1 occupancy between genotypes, including 842 sites with increased MED1 occupancy and 270 sites with decreased MED1 occupancy (Fig. 6B and fig. S10A), demonstrating a general pattern of increased MED1 binding in the absence of MED12. Consistent with a model wherein the kinase module limits access of core Mediator to chromatin, *MED12*-deficient

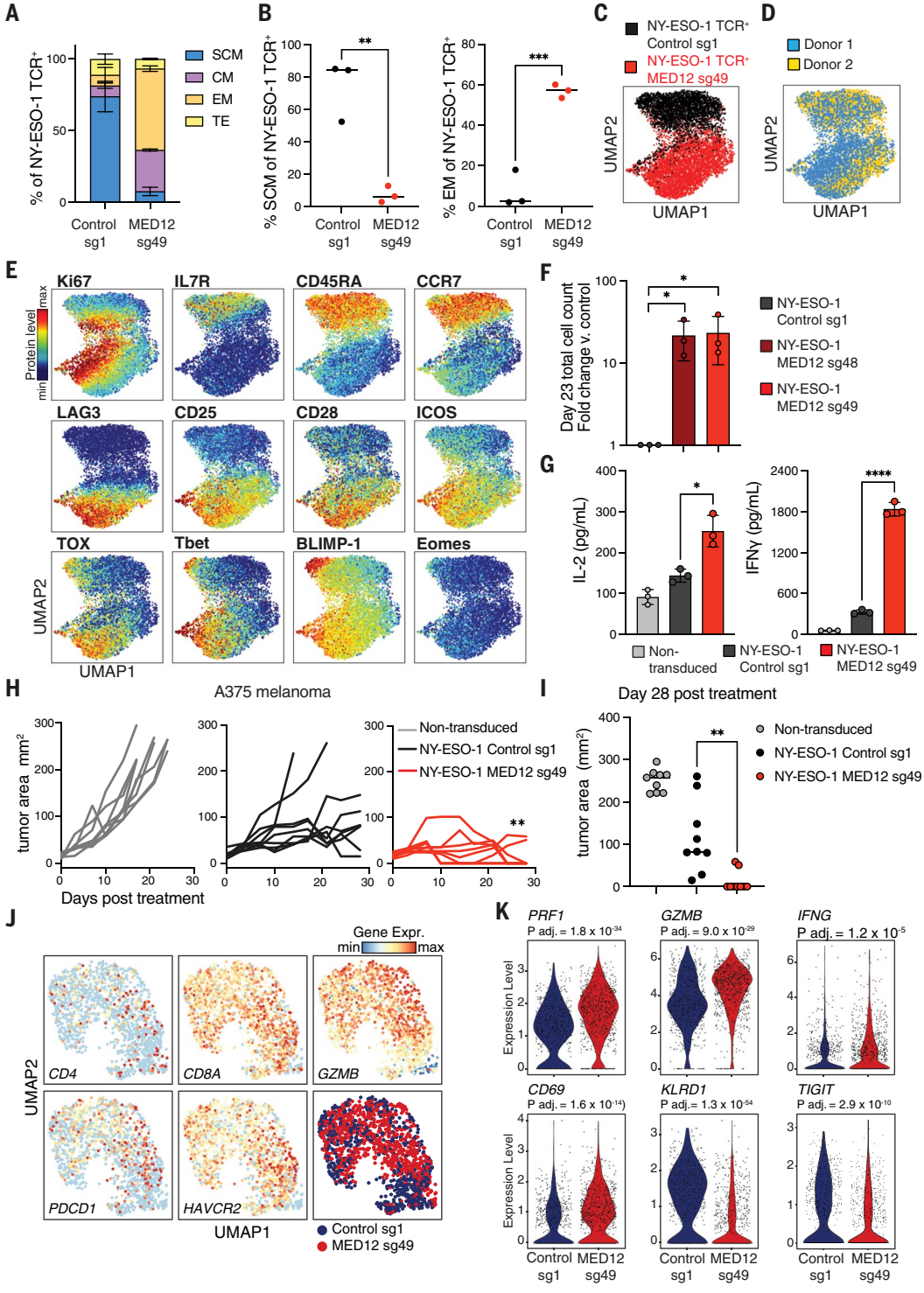


Fig. 4. Loss of *MED12* in T cells using a TCR for tumor recognition enhances antitumor activity and effector function. (**A** and **B**) Frequency of T effector memory cells (CD45RO$^+$, CCR7$^-$) and stem cell memory (CCR7$^+$, CD45RO$^-$) in NY-ESO-1 TCR$^+$ cells 15 day after T cell activation. Data are mean ± SD of n = 3 donors. Two-tailed paired Student's t test. $^{**}P < 0.01$, $^{***}P < 0.001$. SCM, stem central memory; CM, central memory; TE, terminal effector; EM, effector memory. (**C** to **E**) UMAP analysis of control and *MED12*-deficient NY-ESO-1 TCR$^+$ cells. Expression of 34 markers was analyzed by CyTOF (cytometry by time of flight). Control and *MED12*-deficient samples are combined and colored by genotype (C), by donor (D), or by marker intensity (E). Each dot represents a single cell (n = 8319 cells). Data are pooled from two donors. (**F**) In vitro expansion of control and *MED12*-deficient NY-ESO-1 T cells cultured with IL-2. Data are mean ± SD of n = 3 donors. (**G**) IL-2 (left) and IFNγ (right) release after 24-hour coculture of NY-ESO-1 T cells with A375 melanoma cells. Data are mean ± SD from n = 3 cultures. Representative of three independent experiments. (**H** and **I**) Analysis of tumor clearance. Tumor area of NSG mice injected subcutaneously with 3 × 10^6 A375 melanoma cells and treated 7 days later with 3 × 10^6 NY-ESO-1 TCR$^+$ cells. Tumor area was measured by caliper. (H) One-way ANOVA test with Dunnett's multiple comparison test, $^*P < 0.01$. n = 9 mice pooled from two independent experiments. (**J**) scRNA-seq profiles of NY-ESO-1 TCR$^+$ T cells isolated from A375 melanoma tumors by FACS. T cells were administered 12 days after tumor engraftment, and tumors were harvested 6 days after treatment. Each dot represents a single cell and is colored according to gene expression of the indicated genes or by genotype. Data are n = 1542 single cells pooled from four tumors from each genotype. (**K**) Violin plots depicting transcript expression level of selected genes in NY-ESO-1$^+$ tumor infiltrating T cells. n = 669 control sg1 cells and 855 MED12 sg49 cells. Boxes indicate median and interquartile ranges. [(F), (G), and (I)] Two-tailed Student's t test. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

cells showed approximately a twofold increase in MED1 ChIP-seq signal intensity at sites where increased MED1 binding was observed (Fig. 6C). Increased MED1 chromatin binding was confirmed by immunoblotting studies, which demonstrated that, compared with control cells, *MED12*-deficient cells had more abundant MED1 in the chromatin-bound fraction and in total cell lysates, but *MED1* transcript was not differentially expressed, indicating that this effect was posttranscriptionally regulated (Fig. 6, D and E, and fig. S10, B to D).

To determine whether sites with increased MED1 chromatin occupancy after *MED12* deletion represent transcriptionally active regions, we performed ChIP-seq with antibodies targeting H3K27ac and RNAPII pS2 (the elongating form of RNAPII). Consistent with a model

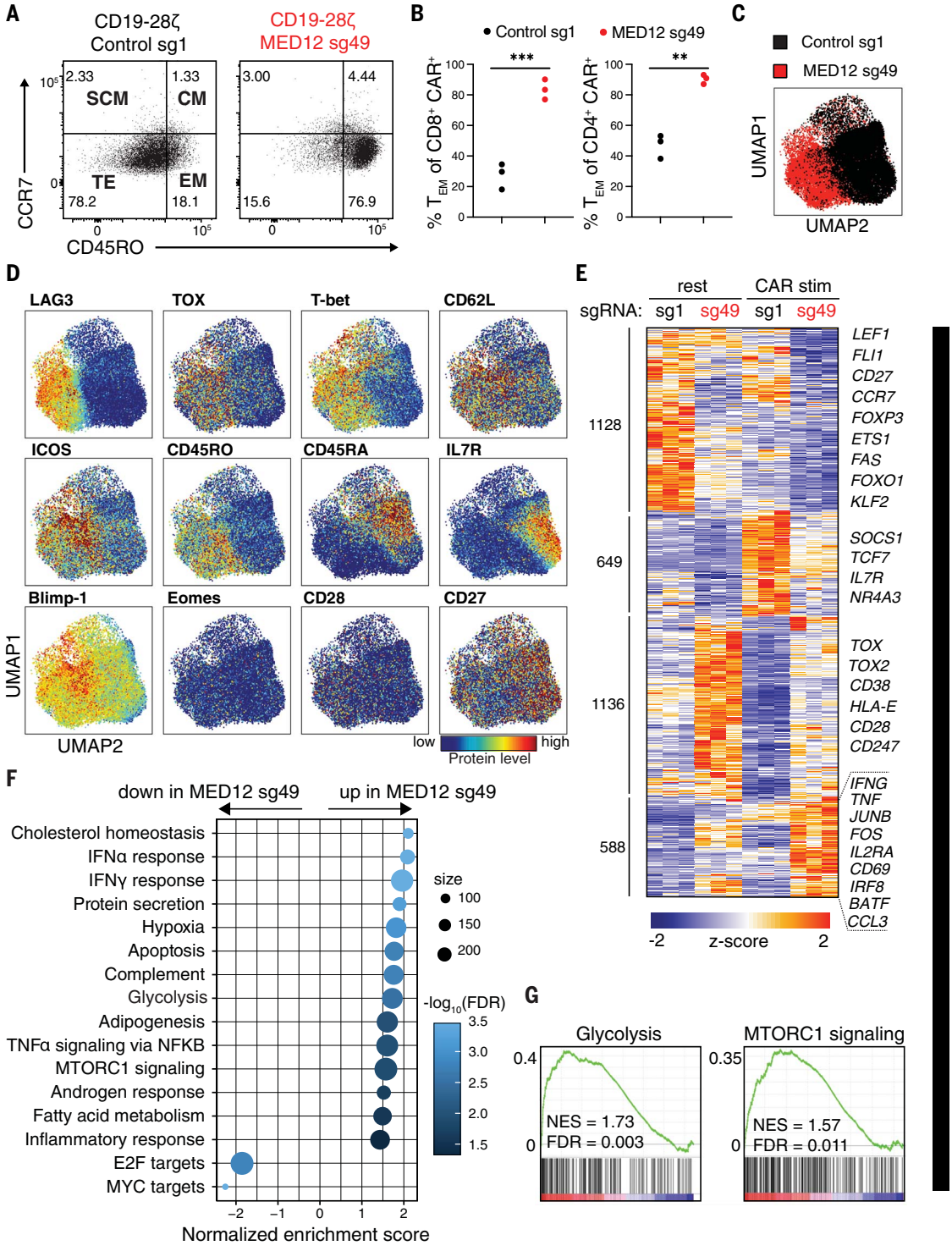


Fig. 5. *MED12*-deficient CD19-28ζ CAR T cells have an effector-like phenotype and an activated transcriptional program. (**A**) Flow cytometry analysis of T cell subsets as assessed by CD45RO and CCR7 expression in control or *MED12*-deficient CD8$^+$ CD19-28ζ CAR T cells 23 days after T cell activation. Representative result of $n = 3$ donors. Gating and subtyping strategy is shown. (**B**) Frequency of T effector memory cells (CD45RO$^+$, CCR7$^-$) in CD8$^+$ and CD4$^+$ CAR T cells. $n = 3$ donors. Two-tailed paired Student's t test. $^{**}P < 0.01$, $^{***}P < 0.001$. (**C** and **D**) UMAP analysis of control and *MED12*-deficient CD19-28ζ CAR T cells 15 days after T cell activation. Expression of 34 markers was analyzed by CyTOF. Control and *MED12*-deficient samples are combined and colored by genotype (C) or by marker intensity (D). Representative donor of $n = 3$ donors. Each dot represents a single cell ($n = 30{,}000$ cells). (**E**) Heatmap of differentially expressed genes in control or *MED12*-deficient CD19-28ζ CAR T cells detected by bulk RNA-seq 15 days after T cell activation. Cells were collected on day 15, and CAR stimulated for 3 hours with plate-bound anti-idiotype antibody. Adjusted $P < 0.01$. $n = 3$ donors. (**F** and **G**) GSEA of unstimulated *MED12*-deficient CAR T cells compared with control cells using the hallmark gene collection. Normalized enrichment scores (NES) and FDR q values are shown. A positive NES indicates that the gene set was enriched in *MED12*-deficient cells.

wherein *MED12* deletion modulates transcription at MED1-bound sites, we observed that sites with changes in MED1 ChIP-seq signal were highly concordant with sites demonstrating differences in H3K27ac ChIP-seq, RNAPII ChIP-seq, and ATAC-seq (assay for transposase-accessible chromatin with sequencing), indicating changed transcriptional activity (Fig. 6F and fig. S10, E and F). Pathway analysis of the genes located near sites with increased MED1 occupancy showed enrichment of genes related to T cell differentiation, cytokine receptor signaling, and immune response genes (Fig. 6G and table S6) and included many genes with significantly increased transcript levels, including *IFNG*, *IL17F*, *IL2RA*, and *TOX* (Fig. 6H and fig. S10G). Together, the data are consistent with a model wherein loss of *MED12* increases MED1 chromatin occupancy at select sites in human T cells responsible for regulating T cell differentiation, leading to transcriptional reprogramming of CAR T cells and enhanced potency.

Loss of MED12 increases STAT5 activity in CD19-28ζ CAR T cells

To define the most differentially regulated transcriptional programs in *MED12*-deficient cells,

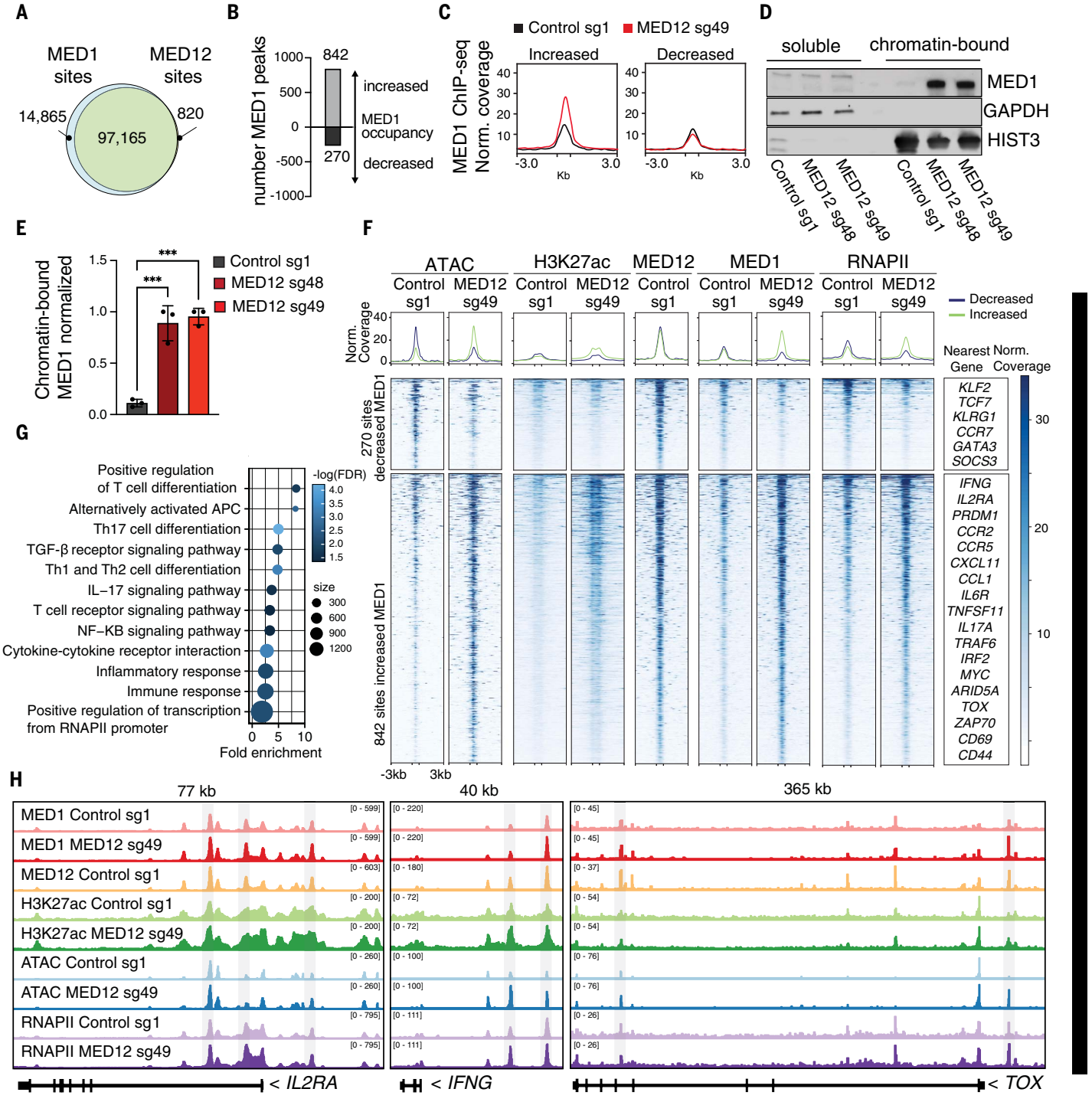


Fig. 6. Loss of *MED12* increases MED1 chromatin occupancy at transcriptionally active enhancers regulating T cell differentiation. (**A**) Venn diagram depicting number of sites bound by MED1 and/or MED12 detected by ChIP-seq in CD19-28ζ control CAR T cells. (**B**) Number of genomic regions with significant change in MED1 occupancy detected by ChIP-seq between *MED12*-deficient and control cells. Adjusted $P < 0.05$. (**C**) Mean normalized ChIP-seq signal at regions with significant differences in MED1 occupancy. (**D**) Western blot analysis of MED1 protein present in soluble and chromatin-bound cellular fractions from control and *MED12*-deficient CD19-28ζ CAR T cells 15 days after T cell activation. Glyceraldehyde phosphate dehydrogenase (GAPDH) and histone 3 (HIST3) are used as markers for each cellular fraction. Representative blot from three independent experiments. (**E**) Densitometric analysis of the Western blot shown in (D). MED1 staining in the chromatin-bound fraction was normalized to HIST3 staining. Data are mean ± SD of $n = 3$ donors. Two-tailed unpaired Student's *t* test. ***$P < 0.001$. (**F**) Heatmap showing ATAC-seq or ChIP-seq coverage at sites with differential MED1 occupancy as defined in (B). (**G**) DAVID (Database for Annotation, Visualization and Integrated Discovery) functional annotation of 842 genes nearest to sites with increased MED1 occupancy in *MED12*-deficient cells compared with control cells. Twelve selected terms of 22 significant terms are shown (FDR < 0.10). Full results table including MED1 ChIP-seq coverage at genes corresponding to Gene Ontology (GO) terms is included in table S6. (**H**) ATAC-seq and MED1, MED12, H3K27ac, and RNAPII ChIP-seq tracks at the *IL2RA*, *IFNG*, and *TOX* loci. [(A) to (C)] Pooled data from $n = 3$ donors. [(F) and (H)] One representative donor of $n = 3$ donors.

we used HOMER motif enrichment analysis to identify differentially accessible transcription factor–binding motifs in *MED12*-deficient versus control cells. The top enriched motifs were STAT5 and STAT1, transcription factors that drive cytokine-mediated gene expression (Fig. 7, A and B, and fig. S11, A to C). AP-1 family motifs including JunB, BATF, and FOS were also significantly enriched, as were motifs from interferon response family (IRF) members such as IRF8 and IRF4. Motif enrichment analysis on sites with increased MED1 occupancy in *MED12*-deficient cells also showed enrichment of STAT and AP-1 motifs (Fig. 7C), confirming that core Mediator is recruited to these sites. Some motifs demonstrated decreased accessibility in *MED12*-deficient cells, including FLI1 and FOXO1, which are implicated in maintaining T cell quiescence (*48*, *49*) (fig. S11D), providing evidence that *MED12* deficiency also diminishes transcription of some genes. ATAC-seq profiling of sorted $CD4^+$ and $CD8^+$ CD19-28ζ CAR T cells showed that chromatin accessibility changes in *MED12*-deficient cells were largely shared between subsets and indicated increased AP-1 and STAT family transcription factor signatures in both $CD4^+$ and $CD8^+$ T cells (fig. S12, A to D).

To assess the functional significance of these findings, we compared STAT5 phosphorylation in *MED12*-deficient versus control cells during in vitro culture with IL-2 and observed that *MED12*-deficient cells manifest significantly higher levels of phosphorylated STAT5 (Fig. 7D). STAT5 activation was dependent on IL-2, given that removal of IL-2 for 24 hours reversed the effect, whereas brief reexposure of *MED12*-deficient CAR T cells to IL-2 led to higher maximum levels of phosphorylated STAT5

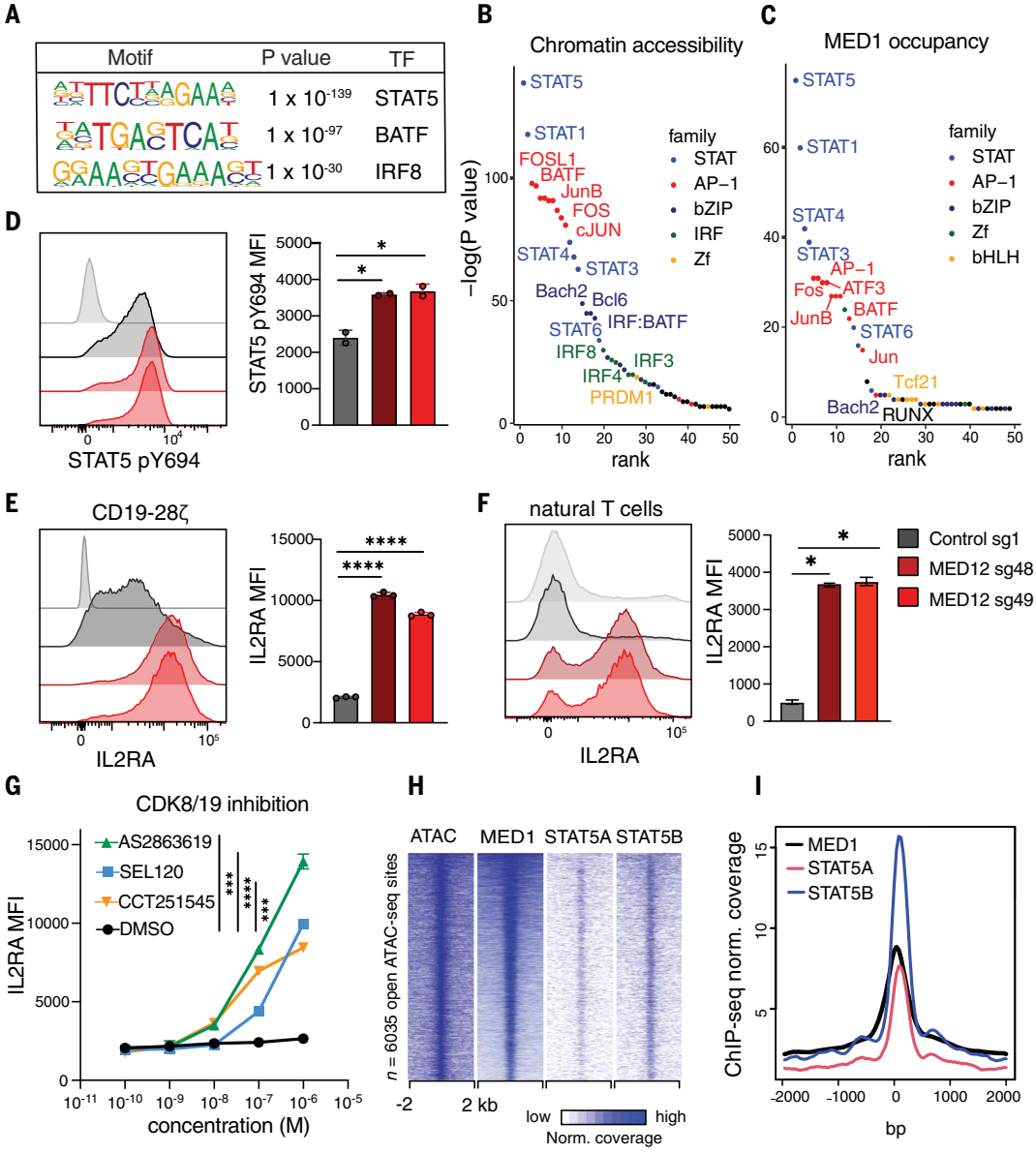


Fig. 7. Loss of *MED12* increases STAT5 activity in CD19-28ζ CAR T cells. (**A** and **B**) Transcription factor binding motif enrichment at sites with increased chromatin accessibility by ATAC-seq in resting *MED12*-deficient CD19-28ζ CAR T cells 15 days after T cell activation. n = 6035 sites with $\log_2$ fold change > 1 and adjusted $P < 0.05$. (**C**) Transcription factor binding motif enrichment at sites with increased MED1 occupancy by ChIP-seq in resting *MED12*-deficient CD19-28ζ CAR T cells. n = 842 sites with adjusted $P < 0.05$. (**D**) Flow cytometry analysis of STAT5 phosphorylation in control and *MED12*-deficient CD19-28ζ CAR T cells 15 days after T cell activation. Cells were cultured in vitro with continuous IL-2. The cells shown in light gray were rested without IL-2 for 24 hours before staining. (**E** and **F**) Flow cytometry analysis of IL2RA expression in control and *MED12*-deficient CD19-28ζ CAR T cells (E) or nontransduced T cells (F) 15 days after activation. Control staining with isotype antibody is shown in light gray. (**G**) Flow cytometry analysis of IL2RA expression in nontransduced T cells cultured with dual inhibitors of CDK8 and CDK19 for 15 days. Inhibitors were supplemented to the culture medium every 48 hours. (**H**) Heatmaps of genomic loci with significantly increased chromatin accessibility by ATAC-seq in *MED12*-deficient CD19-28ζ CAR T cells. Genomic loci are overlaid with ATAC-seq and MED1 ChIP-seq signal from resting CD19-28ζ CAR T cells 15 days after T cell activation. STAT5A and STAT5B ChIP-seq data were obtained from human $CD4^+$ T cells stimulated with IL-2 (GEO accession nos. GSM671400 and GSM671402). (**I**) Mean ChIP-seq signal intensities at n = 6035 sites corresponding to (H). [(A) to (C)] Pooled data from n = 3 donors. HOMER motif enrichment was performed with a set of all [(A) and (B)] ATAC-seq or (C) MED1 ChIP-seq peaks detected in CAR T cells as the background. [(D) to (G)] Data are mean ± SD from n = 2 to 3 wells. Representative results from two independent experiments. Two-tailed unpaired Student's t test. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$. [(H) and (I)] One representative donor of n = 3 donors.

compared with control cells (fig. S13, A and B). *MED12*-deficient CAR T cells also manifested increased expression of the high-affinity IL-2 receptor, *IL2RA*, a downstream target of STAT5-mediated transcription, whereas expression of *IL2RG* was unchanged (Fig. 7E and fig. S13C) (*50*). Furthermore, activated heathy, nontransduced T cells also showed elevated *IL2RA* expression in the absence of *MED12*, indicating that this effect was not dependent on CAR signaling (Fig. 7F), and T cells cultured with CDK8/19 inhibitors also showed elevated expression of *IL2RA*, demonstrating that inhibition of CDK8 and CDK19 kinase activity is sufficient to elicit increased sensitivity to IL-2 in human T cells (Fig. 7G).

We observed increased MED1 occupancy at the *IL2RA* locus in *MED12*-deficient cells (Fig. 6H), and in comparing sites with increased chromatin accessibility in *MED12*-deficient cells to previously published STAT5 ChIP-seq data from $CD4^+$ T cells (*51*), we found extensive colocalization of MED1 and STAT5 (Fig. 7, H and I). Together these results are consistent with a model wherein enhanced MED1 occupancy increases *IL2RA* expression in *MED12*-deficient T cells, thereby initiating a feed-forward loop in which heightened sensitivity to IL-2 promotes STAT5 activation and prolonged upregulation of *IL2RA*. These findings demonstrate that *MED12* deletion regulates T cell effector differentiation and function by transcriptional reprogramming through modulation of MED1 chromatin occupancy. These effects are associated with enhanced activity of multiple transcription factors that control T cell fate during effector differentiation, including STAT5, as well as others such as AP-1, IRF, and other STATs (*52*–*54*).

Discussion

Immune checkpoint inhibitors and adoptive T cell therapies induce profound antitumor effects in some patients for whom all other therapies have failed, but most patients treated with cancer immunotherapies do not experience long-term benefits (*1*–*5*). Enhancing the efficacy of cancer immunotherapies requires new approaches to augment the potency of tumor-specific T cell responses. In the context of adoptive cell therapy, considerable efforts have focused on enhancing stem-like pools of memory T cells and thereby increasing the supply of effector cells (*55*–*57*) and on preventing or reversing T cell exhaustion to enhance T cell function (*13*, *14*, *16*). In this study, we used systematic, unbiased genome-wide CRISPR-Cas9 screens to identify targets capable of enhancing CAR T cell effector function. Our results converged on *MED12* and *CCNC*, core components of the Mediator kinase module, which have not previously been implicated in regulating T cell potency. Our screen also confirmed an expected requisite role for core Mediator in human T cell function.

To explain these findings, we demonstrated that the kinase module and core Mediator are largely colocalized in wild-type CAR T cells, consistent with the known role for the kinase module in regulating the interaction between core Mediator and RNAPII (*58*, *59*), and leading to the hypothesis that loss of *MED12* or *CCNC* in T cells reduces steric hindrance between core Mediator and RNAPII, and thereby increases transcription and modulates T cell function. Consistent with this model, deletion of *MED12* shared some phenotypic similarities with deletion of *CCNC*, although more profound effects were observed after *MED12* deletion, potentially owing to the structural relationships that predict full ablation of the Mediator kinase module in the absence of *MED12*, and partial ablation upon *CCNC* deletion (*30*, *58*, *60*, *61*). Our data demonstrate selective changes in the chromatin landscape and enhancement of MED1 binding to chromatin in *MED12*-deficient T cells, with the most profound effects seen in the regulatory regions associated with transcription factors that regulate T cell differentiation. This selective but widespread pattern of modulation may explain why the Mediator kinase module, rather than any single transcription factor, was the top hit in our CRISPR screens.

MED12 deletion selectively enhanced expression of numerous transcription factors involved in T cell effector differentiation, including TOX, T-bet, and BATF3, and orchestrated coordinated changes in chromatin accessibility and increased MED1 occupancy at motifs for several transcription factor families involved in T cell differentiation, including STAT5, cJUN, BATF, and IRF8. We hypothesize that the effects of *MED12* and *CCNC* deletion are highly context dependent, because we observed increased glycolytic rates in *MED12*- and *CCNC*-deficient effector T cells, consistent with enhanced effector activity, whereas previous studies in cancer have demonstrated diminished glycolytic activity in several cancers following small-molecule mediated inhibition of CDK8 (*62*). The selectivity of the effects is not fully understood, but may be explained in part by the newly reported finding that enhancers have varying degrees of dependence on Mediator for transcriptional activation, determined by the presence of sequence-specific transcription factors and other chromatin characteristics (*63*).

Functional studies revealed elevated STAT5 activity in *MED12*-deficient T cells, manifested as increased *IL2RA* expression and increased sensitivity of T cells to IL-2. We also observed substantial alterations in effector T cell differentiation, with *MED12*-deficient $CD4^+$ and $CD8^+$ cells demonstrating an early activated phenotype, comprising elevated expression of ICOS, Tbet, and CD25, along with simultaneous increases in glycolysis and oxidative phosphorylation (*36*). Paradoxically, however, *MED12*-deficient cells showed elevated CD62L and CD28, molecules not typically expressed on effector cells. This raises the possibility that *MED12* loss induces a synthetic effector state not represented by T cell phenotypes found in nature.

Current models hold that upon antigen encounter, naïve T cells differentiate into $CCR7^-$ $IL7R^-Tbet^+CD62L^-$ SLECs, with most effectors progressing toward a state of terminal differentiation associated with diminished activity, while a fraction express CCR7 and IL7R and differentiate into long-lived memory cells (*44*, *46*, *64*, *65*). The functionally enhanced *MED12*-deficient effector CAR T cells observed here displayed a distinctive $CCR7^-IL7R^-ICOS^+$ $Tbet^+CD62L^+$ phenotype, with diminished expression of CD45RA (*42*). Of interest, $CD4^+$ $ICOS^+Tbet^+LAG3^+$ cells similar to those described here have been implicated in mediating tumor regression in patients treated with anti-CTLA4 therapy (*66*). Although *MED12*-deficient cells showed increases in LAG3 and TOX, which have been associated with T cell exhaustion, we observed down-regulation of CD39 and augmentation of numerous functional attributes spanning improved proliferation, cytokine secretion, and, ultimately, sustained improvements in functionality even at late time points, improved tumor control in vivo, and long-term protection from antigen rechallenge. Together, these data are not consistent with exacerbation of T cell exhaustion in *MED12*-deficient T cells.

This work identifies the Mediator kinase module as a critical negative regulator of T cell effector differentiation and function. Deletion of either *MED12* or *CCNC* induces broad functional enhancements in effector T cells, including increased expansion, cytokine secretion, metabolic fitness, and sustained effector function during chronic stimulation, all properties that would be predicted to enhance antitumor effects. Pharmacological inhibition of Mediator kinase activity phenocopied genetic ablation of *MED12* with regard to increased expansion and elevated expression of *IL2RA*, raising the possibility of synergistic antitumor effects of such agents in the context of immunotherapy, because hyperactive CDK8/19 kinase activity activates oncogenes in some cancer types (*67*). The work further implicates interactions between the kinase module and core Mediator as a major axis of regulation of T cell differentiation. Technologies to inactivate genes in the context of ex vivo cell manufacturing and even in vivo gene editing (*68*), using a variety of approaches, including CRISPR-Cas9, Zinc finger nucleases, TALENs, or base editing, are increasingly available for emerging applications in human medicine (*69*), highlighting the potential for clinical translation of these findings.

Materials and methods

T cell isolation

Whole blood buffy coats were obtained from the Stanford Blood Center from healthy volunteers under 41 years. T cells were isolated using the RosetteSep Human T Cell Enrichment Cocktail (Stemcell Technologies). T cells were stored in CryoStor cell cryopreservation media CS10 (Sigma-Aldrich) in liquid nitrogen.

CRISPR screen

T cell activation and culturing

Two hundred million T cells each from two donors were thawed on day 0 and activated with CD3/CD28 Dynabeads (Invitrogen) at a ratio of three beads per T cell. Cells were cultured in AIM-V medium (Gibco) supplemented with 5% fetal bovine serum (FBS), 10 mM HEPES, 1X penicillin-streptomycin-glutamine supplement (Gibco), and 10 ng/ml recombinant IL-2 (21.8 IU/ml) (Peprotech). Cells were maintained at a density between 0.5 and 2 million per milliliter in T175 flasks.

Lentiviral transduction

The complete Bassik Human CRISPR Knockout Library was obtained from Addgene and amplified with Endura ElectroCompetent Cells (Lucigen). LentiX cells (Takara) were plated on 150-mm plates coated with poly-D-lysine (Corning) and transfected with 18 μg REV, 18 μg GAG/POL, 7 μg VSVg, 15 μg library vector, 3.38 ml Opti-MEM (Gibco), and 135 μl Lipofectamine 2000 (Invitrogen) per plate. Media was changed 24 hours after transfection, and supernatant was harvested 48 hours after transfection. Lentiviral supernatant was concentrated with Lenti-X Concentrator (Takara) and added to the T cell culture medium 2 days after activation. On day 3, cells were assessed for mCherry expression by flow cytometry to confirm that the percentage of transduced cells was between 8 and 12%.

CAS9 electroporation

On day 3, 100 μl reactions were assembled with 10 million T cells, 30 μg Alt-R S.p. Cas9 Nuclease V3 (IDT), 90 μl P3 buffer (Lonza), and 7 μl duplex buffer (IDT). Cells were pulsed with protocol EO115 using the P3 Primary Cell 4D-Nucleofector Kit and 4D-Nucleofector System (Lonza). Cells were recovered immediately with warm media for 6 hours before transduction with CAR. The electroporation protocol was repeated on day 5.

Retroviral transduction

Retroviral supernatant was produced as previously described (*13*). Briefly, 293GP cells were plated on poly-D-lysine (Corning) coated plates and transfected with RD114 envelope and HA-28ζ CAR encoding plasmids using Lipofectamine 2000 (Invitrogen). Media was changed 24 hours after transfection, and supernatant was harvested 48 and 72 hours after transfection. On days 3 and 4, tissue culture plates were coated with RetroNectin (Takara), blocked with 2% bovine serum albumin (BSA) for 5 min, and incubated with retroviral supernatant for 2 hours at 32°C, 3200 rpm. T cells were added to virus-coated plates at a density of 1×10^6 per milliliter. On day 5, CD28/CD3 Dynabeads (Invitrogen) were removed using magnetic separation. Cells were cultured with puromycin at 2.5 μg/ml from days 7 to 10 to eliminate cells that did not express a guide.

Expansion screen

The CAR T cells transduced with the sgRNA library were cultured in T175 flasks and were passaged every other day. On day 15, 100 million Nalm6-GD2 cells were added to 100 million T cells and cocultured to day 23. Fifty percent of the culture volume was discarded at each passage. On day 23, duplicate samples of 30 million cells were collected for genomic DNA extraction. The plasmid DNA encoding the lentiviral sgRNA library was used to approximate the relative abundance of sgRNAs at the start of the experiment.

Cytokine production screen

On day 15, duplicate samples of 30 million cells from each of two donors were harvested from the total population for genomic DNA extraction. One hundred million CAR T cells were cocultured for 6 hours with 100 million Nalm6-GD2 tumor cells with eBioscience Monensin Solution (Invitrogen) in 200 ml medium without IL-2. Intracellular cytokine staining was performed using the Cytofix/Cytoperm Kit (BD). Cells were stained with antibodies specific for CD4, CD8, TNF-α, and IL-2 (table S4) and fixable viability dye eFluor506 (eBioscience). Cell sorting was performed at the Stanford Shared FACS Facility on a FACSAria II equipped with a 70-μm nozzle. The top 10% of $TNF\alpha^+$ and $IL\text{-}2^+$ were sorted using individual gates for $CD4^+$ and $CD8^+$ cells. $CD4^+$ and $CD8^+$ cytokine high cells were pooled. A single sorted sample of ~5 million $TNF\alpha^+$ $IL\text{-}2^+$ cells were collected from each donor for genomic DNA extraction.

Genomic DNA extraction and sequencing library preparation

Technical duplicates were performed for genomic DNA extraction, library preparation, and sequencing. Genomic DNA was extracted from cell pellets using overnight lysis in SDS with proteinase K at 37°C as previously described (*70*). Briefly, protein was precipitated with ammonium acetate, and genomic DNA was precipitated with isopropanol. All the recovered genomic DNA was used as a template for polymerase chain reaction (PCR) to generate the sequencing libraries. The libraries were prepared as previously described (*24*). Illumina sequencing adapters were added using custom primers and sequencing was performed on the Illumina NovaSeq 6000 PE150 platform at a depth of 5×10^7 reads per sample. Sequencing was performed by Novogene (Sacramento, CA).

CRISPR screen data analysis

Guide sequences were extracted from FASTQ files and matched to the Bassik library index using a custom R script. Raw counts for each guide were provided as input to the MAGeCK algorithm (*25*). For the expansion screen, two replicates from plasmid DNA library were compared with four samples collected on day 23 (two from each donor). For the cytokine production screen, four samples collected on day 15 (two from each donor) were compared with two samples (one from each donor) that were sorted for high cytokine expression. The MAGeCK algorithm was used to perform normalization, calculate log fold changes for guides and genes, and calculate adjusted *P* values.

Targeted CRISPR gene editing

Ribonucleoprotein (RNP) was prepared using synthetic sgRNA with 2'-*O*-methyl phosphorothioate modification (Synthego) diluted in TE buffer at 100 μM. Five microliters sgRNA were incubated with 2.5 μl duplex buffer (IDT) and 2.5 μg Alt-R S.p. Cas9 Nuclease V3 (IDT) for 30 min at room temperature. One hundred–microliter reactions were assembled with 10 million T cells, 90 μl P3 buffer (Lonza), and 10 μl RNP. Cells were pulsed with protocol EO115 using the P3 Primary Cell 4D-Nucleofector Kit and 4D-Nucleofector System (Lonza). Cells were recovered immediately with warm media for 6 hours before transduction with CAR. Guides sequences: AAVS1-sg1 5′ GGGGC-CACUAGGGAC-AGGAU 3′, CCNC-sg40 5′ GAUGCCAAAAACA-CACAUGU 3′, CCNC-sg46 5′ GGAUUUAAAGU-UUCUCUCAG 3′, MED12-sg48 5′ CCUGCCU-CAGGAUGAACUGA 3′, and MED12-sg49 5′ UAACCAGCCUGCUGUCUCUG 3′.

Assessment of targeted CRISPR gene editing

Four to seven days after editing, genomic DNA was extracted with QuickExtract DNA Extraction Solution (Lucigen) and ~500–base pair (bp) regions flanking the cut site were amplified with Phusion Hot Start Flex 2X Master Mix (New England Biolabs) according to manufacturer's instructions. Sanger sequencing traces were analyzed by Inference of CRISPR Edits (ICE) (*31*).

Cell lines

Nalm6 leukemia, 143B osteosarcoma and A375 melanoma cells were obtained from American Type Culture Collection. Cell lines were stably transduced with GFP and firefly luciferase. Nalm6-GD2 cells were engineered to stably express GD2 synthase and GD3 synthase to

obtain surface expression of GD2 disialoganglioside. Single-cell clones were selected for high expression of GFP, luciferase, and GD2. Cell lines were maintained in RPMI (Gibco) supplemented with 10 mM HEPES, 10% FBS, and 1X penicillin-streptomycin-glutamine supplement (Gibco).

T cell expansion and viability assays

T cells were activated for 4 days at a 1:3 ratio of T cells to anti-CD3/28 Dynabeads (Invitrogen). T cell expansion assays were performed with IL-2 in the culture medium at 10 ng/ml (21.8 IU/ml) unless indicated otherwise. Cell counts and viability measurements were obtained using the Cellaca Mx Automated Cell Counter (Nexcelom). Cells were stained with acridine orange and propidium iodide to assess viability. A portion of the culture volume was discarded at each passage, and the fraction of cells maintained in culture was used to calculate total cell counts.

CDK8 kinase inhibitor assays

SEL120 (SEL120-34A) hydrochloride, AS2863619, and CCT251545 (Selleckchem) were reconstituted at 5 mM in dimethyl sulfoxide and stored at −80°C. Human primary T cells were plated in 96-well plates with 50,000 T cells per well. Inhibitors were added 24 hours after CD3/CD28 bead activation and were freshly supplemented every 48 hours. CD3/28 beads were removed on day 4 after activation. The reported half maximal inhibitory concentration (IC_{50}) for CDK8 is 4.4 nM, 0.6 nM, and 5 nM for SEL120, AS286319, and CCT251545, respectively. The IC_{50} for CDK19 is 10.4 nM, 4.3 nM, and 6.3 nM for SEL120, AS286319, and CCT251545, respectively.

Serial stimulation assay

Starting from 10 days after activation, CAR T cells were plated at a 1:1 ratio with GFP^+ tumor cells without IL-2. At 48- to 72-hour intervals, cell counts were recorded, and flow cytometry was performed to assess the ratio of T cells to tumor cells. Cocultures were then collected and replated in fresh media and additional tumor cells were added to maintain a 1:1 effector to target ratio.

Cytokine production assays

T cells and tumor cells (5×10^4 of each) were cocultured in 250 μl media without IL-2 in round bottom 96-well plates for 24 hours. Culture supernatants were collected and analyzed by enzyme-linked immunosorbent assay (ELISA). IL-2 and IFNγ were detected with the ELISA MAX kit (Biolegend), and TNFα was detected with the Quantikine kit (R&D Systems). Bead-based multiplex cytokine detection assays were performed at the Human Immune Monitoring Center (Stanford University) using the Luminex platform. Nontransduced T cells, which were CD3/28 activated but not retrovirally transduced or gene-edited, were included as a negative control.

Incucyte cytotoxicity assay

Tumor cells (5×10^4) were cocultured with variable numbers of CAR T cells in 250 μl media without IL-2 in flat bottom 96-well plates for 80 hours. Time lapse microscopy images were obtained with the Incucyte Live Cell Analysis system (Essen Bioscience) at 10× magnification. Total green object integrated intensity (green calibrated units times square micrometers per image) was used to assess tumor killing. Effector to target cell ratios are indicated in the figures or figure legends.

RT-qPCR

RNA was extracted with RNeasy kit (Qiagen), and cDNA was synthesized with High-Capacity cDNA Reverse Transcription kit (ThermoFisher). Reverse transcription–quantitative polymerase chain reaction (RT-qPCR) was performed with PowerUp SYBR Green Master Mix (ThermoFisher) using the Bio-Rad CFX thermocycler and CFX Manager software. Target gene expression was normalized to the 18*S* housekeeping gene using the 2ΔΔCt method. Primer sequences for RT-PCR: IFNG-F 5′ TGACCAGAGCATCCAAAAGA 3′, IFNG-R 5′ CTCTTCGACCTCGAAACAGC 3′, IL2-F 5′ TGCATTGCACTAAGTCTTGCAC 3′, IL2-R 5′ AGTTCTGTGGCCTTCTTGGG 3′, TNF-F 5′ CACAGTGAAGTGCTGGCAAC 3′, TNF-R 5′ AGGAAGGCCTAAGGTCCACT 3′, 18S-F 5′ GCAGAATCCACGCCAGTACAAG 3′, 18S-R 5′ GCTTGTTGTCCAGACCATTGG 3′.

Seahorse assay

Metabolic analysis was carried out using Seahorse Bioscience Analyzer XFe96. Briefly, 2×10^6 cells were resuspended in extracellular flux (XF) assay media supplemented with 25 mM glucose, 2 mM glutamine, and 1 mM sodium pyruvate and plated on a Cell-Tak (Corning)–coated microplate allowing the adhesion of CAR T cells. Mitochondrial stress and glycolytic parameters were measured by the oxygen consumption rate (OCR) (pmol/min) and extracellular acidification rate (ECAR) (mpH/min), respectively, with use of real-time injections of oligomycin (1.5 μM), carbonyl cyanide *p*-trifluoromethoxyphenylhydrazone (FCCP; 1 μM), and rotenone and antimycin (both 1 μM). Respiratory parameters were calculated according to manufacturer's instructions (Seahorse Bioscience). All chemicals were purchased from Agilent unless stated otherwise.

Flow cytometry

T cells were washed in FACS buffer [Dulbecco's phosphate-buffered saline (DPBS) no calcium, no magnesium (Gibco) with 2% FBS]. Cells were incubated on ice in FACS buffer with antibodies specific for cell surface markers for 20 min. Antibodies used are listed in table S4. For pSTAT5 staining, cells were prepared with the Fix and Perm Cell Permeabilization kit (ThermoFisher) according to manufacturer's instructions. For pS6 staining, T cells were cultured 24 hours in complete AIM-V media without IL-2 prior anti-idiotype stimulation (1 μg/ml of mouse immunoglobulin G anti-human CD19 idiotype cross-linked with 10 μg/ml of anti-mouse Fab in phosphate-buffered saline (PBS). Cells were incubated with anti-idiotype—or left unstimulated—for 1 hour at 37°C and kept on ice immediately afterwards. Cells were fixed (BD Phosphoflow Fix Buffer I) and then permeabilized (BD Perm Buffer III) according to manufacturer's instructions. For mitochondrial mass staining, cells were stained with MitoTracker Green (Cell Signaling Technology) at 200 nM, 37°C for 30 min. Before acquisition, cells were resuspended in FACS buffer and analyzed on a LSRFortessa (BD) with BD FACSDiva software.

Western blotting

Total cell lysates were extracted in nondenaturing lysis buffer (150 mM NaCl, 50 mM Tris pH 8, 1% NP-40, 0.25% sodium deoxycholate with Halt Protease Inhibitor Cocktail) (ThermoFisher Scientific). Chromatin-bound and soluble cellular fractions were prepared with cytoskeletal lysis buffer [10 mM PIPES-KOH (pH 6.8), 100 mM NaCl, 300 mM sucrose, 3 mM MgCl2], and Halt Protease Inhibitor Cocktail. Briefly, cells were washed in PBS, resuspended in lysis buffer, and incubated on ice for 20 min. Cells were centrifuged at 5000 rpm to separate the soluble and chromatin-bound fractions. The soluble fraction was cleared by centrifugation at 13,000 rpm. The chromatin-bound fraction was resuspended in Sample Reducing Buffer (Pierce) and incubated at 100°C for 5 min. Protein concentration was assessed with the DC Protein Assay kit (Bio-Rad), and 20 μg total protein was loaded per sample. Equal volumes of soluble and chromatin fractions were loaded for each sample. SDS–polyacrylamide gel electrophoresis was performed, and proteins were transferred to polyvinylidene difluoride membranes for immunoblotting. Antibodies used are listed in table S4. Immunofluorescence was detected with the Odyssey Imaging System (Licor), or chemiluminescence was detected with autoradiography film.

Western blot quantification

Images captured with autoradiography film were scanned at 600 dpi in 16-bit gray scale, and images captured with the Odyssey Imaging System were exported as JPEGs. Quantification was performed with ImageJ software. A region of interest (ROI) of equal size was used to measure the specific band and background signal in each lane. Pixel densities

were subtracted from 255 to invert the image, and the background values were subtracted from band values to adjust for background signal. For each sample, the background adjusted MED1 pixel density was divided by the same value from the HIST3 loading control to calculate a ratio of MED1 to HIST3. Ratios from donor 1 and donor 2 were normalized to the largest ratio collected in each independent experiment.

Mice

Immunocompromised NOD *scid* IL2Rgamma^{null} (NSG, Strain #005557) mice were purchased from JAX and bred in-house under sterile conditions. Mice were monitored daily by the Veterinary Service Center staff. Care and treatments were in compliance with Stanford University APLAC protocols. Leukemia cells and CAR T cells were administered via intravenous injection. 143B osteosarcoma cells were administered by intramuscular injection. For some experiments, tumor burden was assessed before treatment and mice were randomized to groups to ensure equal tumor burden between treatment groups. Time of treatment and dosing is indicated in figure legends. Researchers were blinded during administration of T cells. Leukemia progression was monitored using the Lago SII (Spectral Instruments Imaging). Quantification of bioluminescence was performed with Aura software (Spectral Instruments Imaging). Solid tumor progression was followed using caliper measurements of the injected leg area. Researchers were blinded to the treatment groups during solid tumor measurements. Mice were euthanized upon manifestation of paralysis, impaired mobility, poor body condition (score of BC2−), or when tumor diameter exceed 17 mm. Sample sizes of five mice per group were selected on the basis of previous experience with these models. All experiments were repeated twice with different donors, and donors used for in vivo experiments were different than those in the screening experiments.

Blood analysis

Blood was collected from the retro-orbital sinus into Microvette blood collection tubes with EDTA (Fisher Scientific). Whole blood was labeled with anti-CD45 (HI30, ThermoFisher), and red blood cells were lysed with FACS Lysing Solution (BD) according to manufacturer's instructions. Samples were mixed with CountBright Absolute Counting beads (ThermoFisher) before flow cytometry analysis.

CyTOF sample preparation and data analysis

CAR T cells (2×10^6) were washed in PBS and resuspended in 250 nM cisplatin (Fluidigm) for 3 min. Cells were washed twice in cell staining medium (CSM; PBS with 0.05% BSA and 0.02% sodium azide) followed by fixing in 1.6% paraformaldehyde for 10 min at room temperature. Cells were washed twice in PBS, flash frozen on dry ice, and stored at −80°C. Cells were thawed, washed in CSM, and barcoded with the Cell-ID 20-plex kit (Fluidigm) as previously described (*71*). Samples were pooled and stained for cell surface markers for 30 min at room temperature. Panel of antibodies can be found in table S3. CARs were detected with anti-idiotype antibodies conjugated to metals. The NY-ESO-1 TCR was detected with PE labeled tetramers, followed by 30 min of secondary staining with anti-PE-156Gd (Fluidigm). Intracellular staining was performed using permeabilization buffer (eBioscience) for 45 min on ice followed by two CSM washes. Cells were resuspended with Cell-ID Intercalator-ID (Fluidigm), washed twice in deionized water, resuspended in 1X EQ beads, and acquired on a Helios mass cytometer (Fluidigm). After acquisition, data were normalized using MATLAB-based software (*72*) and debarcoded using MATLAB-debarcoder. Fsc files were uploaded to the OMIQ platform for analysis (OMIQ.ai).

Bulk RNA-seq

CAR T cells were collected on day 15 and processed without freezing. RNA was extracted using the RNeasy mini kit (Qiagen). RNA quality was assessed by BioAnalyzer (Agilent). Sequencing libraries were prepared by Novogene (Sacramento, CA), and 150 bp paired-end sequencing at a depth of 3×10^7 reads per sample was obtained using the Illumina NovaSeq6000 platform. FASTQ files were generated by Novogene. Transcripts were quantified with Salmon, and DESeq2 was used to identify differentially expressed genes. Gene set enrichment analysis was performed using GSEA software (Broad Institute).

ChIP-seq

ChIP-seq was performed as previously described with minor modifications (*73*). CAR T cells (3×10^6) were double cross-linked by 50 mM disuccinimidyl glutarate (DSG, #C1104 - ProteoChem) for 30 min followed by 10 min of 1% formaldehyde. Formaldehyde was quenched by the addition of glycine. Nuclei were isolated with ChIP lysis buffer (1% Triton x-100, 0.1% SDS, 150 mM NaCl, 1mM EDTA, and 20 mM Tris, pH 8.0). Nuclei were sheared with Covaris sonicator using the following setup: Fill level – 10, Duty Cycle – 5, PIP – 140, Cycles/Burst – 200, Time – 4 min. Sheared chromatin was immunoprecipitated overnight with the antibodies shown in table S4. Antibody chromatin complexes were pulled down with Protein A magnetic beads and washed once in IP wash buffer I [1% Triton, 0.1% SDS, 150 mM NaCl, 1 mM EDTA, 20 mM Tris, pH 8.0, and 0.1% sodium deoxycholate (NaDOC)], twice in IP wash buffer II (1% Triton, 0.1% SDS, 500 mM NaCl, 1 mM EDTA, 20 mM Tris, pH 8.0, and 0.1% NaDOC), once in IP wash buffer III (0.25 M LiCl, 0.5% NP-40, 1 mM EDTA, 20 mM Tris, pH 8.0, 0.5% NaDOC), and once in TE buffer (10 mM EDTA and 200 mM Tris, pH 8.0). DNA was eluted from the beads by vigorous shaking for 20 min in elution buffer (100 mM $NaHCO_3$, 1% SDS). DNA was decross-linked overnight at 65°C and purified with MinElute PCR purification kit (Qiagen). DNA was quantified by Qubit, and 10 ng DNA was used for sequencing library construction with the Ovation Ultralow Library System V2 (Tecan) using 12 PCR cycles. Libraries were sequenced on the Illumina NovaSeq 6000 PE150 platform at a depth of 3×10^7 reads per sample.

ATAC-seq

Approximately 100,000 CAR T cells were used per sample. Nuclei were isolated with ATAC-LB (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% IGEPAL) and used for tagmentation using Nextera DNA Library Preparation Kit (Illumina) from three donors. After tagmentation DNA was purified with MinElute PCR Purification Kit (Qiagen). Tagmented DNA was then amplified with Phusion high-fidelity PCR master mix (NEB) using 14 PCR cycles. Amplified libraries were purified again with MinElute PCR Purification Kit. Fragment distribution of libraries was assessed with Agilent Bioanalyzer and libraries were sequenced on the Illumina NovaSeq 6000 PE150 platform at a depth of 3×10^7 reads per sample. Sequencing was performed by Novogene (Sacramento, CA).

ATAC-seq and ChIP-seq data processing

Quality control, aligning, and signal tracks

Paired end FASTQ files were trimmed to remove adapters and low-quality sequences using "fastp" and then were aligned to the hg38 reference genome using "hisat2" with the "–no-spliced-alignment" and "–very-sensitive" options. Duplicates were marked and removed with "picard MarkDuplicates." Reads with high-quality concordant alignments to human chromosomes chr1-chr22 and chrX (e.g., excluding chrY and chrM) were converted to BED files for downstream processing. Bigwig files were normalized by using the number of fragments overlapping peaks genome wide, which controls for differences in sequencing depth and also library quality between samples, were exported from R using "rtracklayer::export" and visualized using the Integrative Genomics Viewer (IGV). Plots were generated in R using ChIPpeakAnno (3.13) (*74*).

Obtaining a union peak set and counts matrix

Peaks were called for each replicate using MACS2. Reproducible peaks for each sample were determined as peaks present in at least two of three replicates and merged to create a

disjoint peak set for each sample. Reproducible peaks for each sample were then merged into a disjoint union peak set encompassing all samples using our previously described iterative procedure which merges the peak sets by repeatedly removing less significant overlapping peaks until no overlaps remain. The peak by sample counts matrix was obtained by counting reads overlapping with each peak. The counts matrix was loaded into DESeq2 for differential analysis. Selected peak sets were also visualized across samples by using the bigwig track files described above with "plotHeatmap" from the "deepTools" analysis suite.

Details of peak sets

Differential "up" and "down" peak sets were obtained from DESeq2 pairwise comparisons using false discovery rate (FDR) ≥ 0.05 and sometimes an additional fold change threshold, as indicated. "Bound" peak sets: for each sample (where a sample is, for example, "h3k27ac_MED12_unstim"), bound peaks were determined on the basis of thresholding the normalized counts matrix. The normalized counts matrix was obtained by multiplying each column (replicate) in the counts matrix by (1×10^6 per total reads in peaks in that replicate). The bound peaks for each sample were then determined as peaks that had at least two of three replicates with a normalized value greater than a threshold. A cutoff of 2 was selected empirically, which yielded between 60,000 and 120,000 peaks per sample. Finally, bound peak sets for each sample were merged into bound peak sets for each sample set (e.g., an overall h3k27ac CAR T peak set) by taking the union.

Motif analysis

Motifs enriched in particular peak sets were analyzed using the HOMER "findMotifsGenome.pl" utility or chromVAR for ATAC-seq samples. For HOMER analysis, total bound peaks for each sample set were used as background, thus, the enrichments represent motif enrichments relative to general CAR T–specific peaks rather than enrichments relative to the whole genome.

Single-cell RNA-seq

Isolation of tumor infiltrating NY-ESO-1$^+$ T cells

Tumors were harvested 6 days after adoptive transfer of NY-ESO-1$^+$ T cells. Tissue dissociation was performed with the Mouse Tumor Dissociation Kit (Miltenyi) and Gentle MACS C tubes (Miltenyi) according to manufacturer's instructions. Red blood cells were lysed (Red Blood Cell Lysis Solution, Miltenyi), and samples were washed and passed through 70-μm filters before staining for 30 min at room temperature with fixable viability dye eFluor506 (eBioscience), antibodies specific for Vβ13.1, the beta chain of the NY-ESO-1 TCR (table S4), and tetramers loaded with the SLLMWITQC peptide. Cell sorting of live, tetramer$^+$, Vβ13.1$^+$ cells was performed at the Stanford Shared FACS Facility on a FACSAria II equipped with a 70-μm nozzle.

scRNA-seq library preparation and sequencing

Cells harvested from four tumors from each experimental condition were pooled and capture was performed at the Stanford Function Genomics Facility using the 10X Genomics platform. Sequencing libraries were prepared with the 10X Chromium Next GEM Single Cell 3′ v 3.1 kit (10X Genomics) according to manufacturer's instructions. Deep sequencing of single-index libraries was performed by Novogene (Sacramento, CA) on the Illumina NovaSeq 6000 PE150 platform with an average depth of ~38,000 reads per cell.

scRNA-seq data processing and analysis

Index demultiplexing and generation of FASTQ files was performed by Novogene. Alignment to hg38 was performed with Cell Ranger v7.0.0 (10X Genomics). Filtering was performed to remove debris and dead cells using the Seurat package v4.1.1 in R. Clustering of scRNA-seq profiles, dimensionality reduction using uniform manifold approximation and projection (UMAP), and differential gene analysis were performed with Seurat. For comparison to published datasets, marker genes for each cluster of Zheng *et al.* were obtained from table S3 of the Zheng *et al.* manuscript (*41*). For each cluster, the top 50 marker genes were obtained by using the effect size ranking (column "comb.ES.rank" in Zheng *et al.* table S3). A module score for each gene set was added to each cell in our TIL scRNA-seq dataset with the Seurat function "AddModuleScore" and then visualized by genotype (MED12-sg49 versus control sg1). Statistical significance was assessed via two-sided Wilcoxon test using R function "wilcox.test."

REFERENCES AND NOTES

1. S. L. Maude *et al.*, Tisagenlecleucel in children and young adults with B-cell lymphoblastic leukemia. *N. Engl. J. Med.* **378**, 439–448 (2018). doi: 10.1056/NEJMoa1709866; pmid: 29385370
2. D. W. Lee *et al.*, T cells expressing CD19 chimeric antigen receptors for acute lymphoblastic leukaemia in children and young adults: A phase 1 dose-escalation trial. *Lancet* **385**, 517–528 (2015). doi: 10.1016/S0140-6736(14)61403-3; pmid: 25319501
3. A. Ribas, J. D. Wolchok, Cancer immunotherapy using checkpoint blockade. *Science* **359**, 1350–1355 (2018). doi: 10.1126/science.aar4060; pmid: 29567705
4. R. G. Majzner, C. L. Mackall, Clinical lessons learned from the first leg of the CAR T cell journey. *Nat. Med.* **25**, 1341–1355 (2019). doi: 10.1038/s41591-019-0564-6; pmid: 31501612
5. F. L. Locke *et al.*, Long-term safety and activity of axicabtagene ciloleucel in refractory large B-cell lymphoma (ZUMA-1): A single-arm, multicentre, phase 1-2 trial. *Lancet Oncol.* **20**, 31–42 (2019). doi: 10.1016/S1470-2045(18)30864-7; pmid: 30518502
6. S. A. Rosenberg *et al.*, Treatment of patients with metastatic melanoma with autologous tumor-infiltrating lymphocytes and interleukin 2. *J. Natl. Cancer Inst.* **86**, 1159–1166 (1994). doi: 10.1093/jnci/86.15.1159; pmid: 8028037
7. P. F. Robbins *et al.*, Tumor regression in patients with metastatic synovial cell sarcoma and melanoma using genetically engineered lymphocytes reactive with NY-ESO-1. *J. Clin. Oncol.* **29**, 917–924 (2011). doi: 10.1200/JCO.2010.32.2537; pmid: 21282551
8. S. P. D'Angelo *et al.*, Antitumor activity associated with prolonged persistence of adoptively transferred NY-ESO-1^{c259} T cells in synovial sarcoma. *Cancer Discov.* **8**, 944–957 (2018). doi: 10.1158/2159-8290.CD-17-1417; pmid: 29891538
9. J. A. Fraietta *et al.*, Determinants of response and resistance to CD19 chimeric antigen receptor (CAR) T cell therapy of chronic lymphocytic leukemia. *Nat. Med.* **24**, 563–571 (2018). doi: 10.1038/s41591-018-0010-1; pmid: 29713085
10. M. Philip *et al.*, Chromatin states define tumour-specific T cell dysfunction and reprogramming. *Nature* **545**, 452–456 (2017). doi: 10.1038/nature22367; pmid: 28514453
11. S. Srivastava *et al.*, Immunogenic chemotherapy enhances recruitment of CAR-T cells to lung tumors and improves antitumor efficacy when combined with checkpoint blockade. *Cancer Cell* **39**, 193–208.e10 (2021). doi: 10.1016/j.ccell.2020.11.005; pmid: 33357452
12. K. G. Anderson, I. M. Stromnes, P. D. Greenberg, Obstacles posed by the tumor microenvironment to T cell activity: A case for synergistic therapies. *Cancer Cell* **31**, 311–325 (2017). doi: 10.1016/j.ccell.2017.02.008; pmid: 28292435
13. R. C. Lynn *et al.*, c-Jun overexpression in CAR T cells induces exhaustion resistance. *Nature* **576**, 293–300 (2019). doi: 10.1038/s41586-019-1805-z; pmid: 31802004
14. J. Chen *et al.*, NR4A transcription factors limit CAR T cell function in solid tumours. *Nature* **567**, 530–534 (2019). doi: 10.1038/s41586-019-0985-x; pmid: 30814732
15. L. Labanieh *et al.*, Enhanced safety and efficacy of protease-regulated CAR-T cell receptors. *Cell* **185**, 1745–1763.e22 (2022). doi: 10.1016/j.cell.2022.03.041; pmid: 35483375
16. E. W. Weber *et al.*, Transient rest restores functionality in exhausted CAR-T cells through epigenetic remodeling. *Science* **372**, eaba1786 (2021). doi: 10.1126/science.aba1786; pmid: 33795428
17. E. A. Stadtmauer *et al.*, CRISPR-engineered T cells in patients with refractory cancer. *Science* **367**, eaba7365 (2020). doi: 10.1126/science.aba7365; pmid: 32029687
18. E. Shifrut *et al.*, Genome-wide CRISPR screens in primary human T cells reveal key regulators of immune function. *Cell* **175**, 1958–1971.e15 (2018). doi: 10.1016/j.cell.2018.10.024; pmid: 30449619
19. D. Wang *et al.*, CRISPR screening of CAR T cells and cancer stem cells reveals critical dependencies for cell-based therapies. *Cancer Discov.* **11**, 1192–1211 (2021). doi: 10.1158/2159-8290.CD-20-1243; pmid: 33328215
20. J. A. Belk *et al.*, Genome-wide CRISPR screens of T cell exhaustion identify chromatin remodeling factors that limit T cell persistence. *Cancer Cell* **40**, 768–786.e7 (2022). doi: 10.1016/j.ccell.2022.06.001; pmid: 35750052
21. L. El Khattabi *et al.*, A pliable Mediator acts as a functional rather than an architectural bridge between promoters and enhancers. *Cell* **178**, 1145–1158.e20 (2019). doi: 10.1016/j.cell.2019.07.011; pmid: 31402173
22. W. A. Whyte *et al.*, Master transcription factors and Mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013). doi: 10.1016/j.cell.2013.03.035; pmid: 23582322
23. M. Quevedo *et al.*, Mediator complex interaction partners organize the transcriptional network that defines neural stem cells. *Nat. Commun.* **10**, 2669 (2019). doi: 10.1038/s41467-019-10502-8; pmid: 31209209
24. D. W. Morgens *et al.*, Genome-scale measurement of off-target activity using Cas9 toxicity in high-throughput screens. *Nat. Commun.* **8**, 15178 (2017). doi: 10.1038/ncomms15178; pmid: 28474669
25. W. Li *et al.*, MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.* **15**, 554 (2014). doi: 10.1186/s13059-014-0554-4; pmid: 25476604
26. L. Sabbagh *et al.*, The selective increase in caspase-3 expression in effector but not memory T cells allows susceptibility to apoptosis. *J. Immunol.* **173**, 5425–5433 (2004). doi: 10.4049/jimmunol.173.9.5425; pmid: 15494489
27. R. G. Majzner *et al.*, Tuning the antigen density requirement for CAR T-cell activity. *Cancer Discov.* **10**, 702–723 (2020). doi: 10.1158/2159-8290.CD-19-0945; pmid: 32193224

28. A. Verger, D. Monté, V. Villeret, Twenty years of Mediator complex structural studies. *Biochem. Soc. Trans.* **47**, 399–410 (2019). doi: 10.1042/BST20180608; pmid: 30733343
29. R. Abdella *et al*., Structure of the human Mediator-bound transcription preinitiation complex. *Science* **372**, 52–56 (2021). doi: 10.1126/science.abg3074; pmid: 33707221
30. Y. C. Li *et al*., Structure and noncanonical Cdk8 activation mechanism within an Argonaute-containing Mediator kinase module. *Sci. Adv.* **7**, eabd4484 (2021). doi: 10.1126/sciadv.abd4484; pmid: 33523904
31. D. Conant *et al*., Inference of CRISPR Edits from Sanger Trace Data. *CRISPR J.* **5**, 123–130 (2022). doi: 10.1089/crispr.2021.0113; pmid: 35119294
32. A. H. Long *et al*., 4-1BB costimulation ameliorates T cell exhaustion induced by tonic signaling of chimeric antigen receptors. *Nat. Med.* **21**, 581–590 (2015). doi: 10.1038/nm.3838; pmid: 25939063
33. N. Li *et al*., Cyclin C is a haploinsufficient tumour suppressor. *Nat. Cell Biol.* **16**, 1080–1091 (2014). doi: 10.1038/ncb3046; pmid: 25344755
34. M. B. Moyo, J. B. Parker, D. Chakravarti, Altered chromatin landscape and enhancer engagement underlie transcriptional dysregulation in MED12 mutant uterine leiomyomas. *Nat. Commun.* **11**, 1019 (2020). doi: 10.1038/s41467-020-14701-6; pmid: 32094355
35. R. I. Klein Geltink, R. L. Kyle, E. L. Pearce, Unraveling the complex interplay between T cell metabolism and function. *Annu. Rev. Immunol.* **36**, 461–488 (2018). doi: 10.1146/annurev-immunol-042617-053019; pmid: 29677474
36. L. S. Levine *et al*., Single-cell analysis by mass cytometry reveals metabolic states of early-activated CD8$^+$ T cells during the primary immune response. *Immunity* **54**, 829–844.e5 (2021). doi: 10.1016/j.immuni.2021.02.018; pmid: 33705706
37. S. Krishna *et al*., Stem-like CD8 T cells mediate response of adoptive cell immunotherapy against human cancer. *Science* **370**, 1328–1334 (2020). doi: 10.1126/science.abb9847; pmid: 33303615
38. P. K. Gupta *et al*., CD39 expression identifies terminally exhausted CD8$^+$ T cells. *PLOS Pathog.* **11**, e1005177 (2015). doi: 10.1371/journal.ppat.1005177; pmid: 26485519
39. R. Thomas *et al*., NY-ESO-1 based immunotherapy of cancer: Current perspectives. *Front. Immunol.* **9**, 947 (2018). doi: 10.3389/fimmu.2018.00947; pmid: 29770138
40. C. R. Good *et al*., An NK-like CAR T cell transition in CAR T cell dysfunction. *Cell* **184**, 6081–6100.e26 (2021). doi: 10.1016/j.cell.2021.11.016; pmid: 34861191
41. L. Zheng *et al*., Pan-cancer single-cell landscape of tumor-infiltrating T cells. *Science* **374**, eabe6474 (2021). doi: 10.1126/science.abe6474; pmid: 34914499
42. D. Hamann *et al*., Phenotypic and functional separation of memory and effector human CD8$^+$ T cells. *J. Exp. Med.* **186**, 1407–1418 (1997). doi: 10.1084/jem.186.9.1407; pmid: 9348298
43. P. Romero *et al*., Four functionally distinct populations of human effector-memory CD8$^+$ T lymphocytes. *J. Immunol.* **178**, 4112–4119 (2007). doi: 10.4049/jimmunol.178.7.4112; pmid: 17371966
44. F. Sallusto, D. Lenig, R. Förster, M. Lipp, A. Lanzavecchia, Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. *Nature* **401**, 708–712 (1999). doi: 10.1038/44385; pmid: 10537110
45. Y. Muroyama, E. J. Wherry, Memory T-cell heterogeneity and terminology. *Cold Spring Harb. Perspect. Biol.* **13**, a037929 (2021). doi: 10.1101/cshperspect.a037929; pmid: 33782027
46. N. S. Joshi *et al*., Inflammation directs memory precursor and short-lived effector CD8$^+$ T cell fates via the graded expression of T-bet transcription factor. *Immunity* **27**, 281–295 (2007). doi: 10.1016/j.immuni.2007.07.010; pmid: 17723218
47. J. Soutourina, Transcription regulation by the Mediator complex. *Nat. Rev. Mol. Cell Biol.* **19**, 262–274 (2018). doi: 10.1038/nrm.2017.115; pmid: 29209056
48. Z. Chen *et al*., *In vivo* CD8$^+$ T cell CRISPR screening reveals control by Fli1 in infection and cancer. *Cell* **184**, 1262–1280.e22 (2021). doi: 10.1016/j.cell.2021.02.019; pmid: 33636129
49. A. Delpoux *et al*., FOXO1 constrains activation and regulates senescence in CD8 T cells. *Cell Rep.* **34**, 108674 (2021). doi: 10.1016/j.celrep.2020.108674; pmid: 33503413
50. H. P. Kim, J. Imbert, W. J. Leonard, Both integrated and differential regulation of components of the IL-2/IL-2 receptor system. *Cytokine Growth Factor Rev.* **17**, 349–366 (2006). doi: 10.1016/j.cytogfr.2006.07.003; pmid: 16911870
51. W. Liao, J.-X. Lin, L. Wang, P. Li, W. J. Leonard, Modulation of cytokine receptors by IL-2 broadly regulates differentiation into helper T cell lineages. *Nat. Immunol.* **12**, 551–559 (2011). doi: 10.1038/ni.2030; pmid: 21516110
52. P. Tripathi *et al*., STAT5 is critical to maintain effector CD8$^+$ T cell responses. *J. Immunol.* **185**, 2116–2124 (2010). doi: 10.4049/jimmunol.1000842; pmid: 20644163
53. T. W. Hand *et al*., Differential effects of STAT5 and PI3K/AKT signaling on effector and memory CD8 T-cell survival. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 16601–16606 (2010). doi: 10.1073/pnas.1003457107; pmid: 20823247
54. G. Verdeil, D. Puthier, C. Nguyen, A.-M. Schmitt-Verhulst, N. Auphan-Anezin, STAT5-mediated signals sustain a TCR-initiated gene expression program toward differentiation of CD8 T cell effectors. *J. Immunol.* **176**, 4834–4842 (2006). doi: 10.4049/jimmunol.176.8.4834; pmid: 16585578
55. C. S. Jansen *et al*., An intra-tumoral niche maintains and differentiates stem-like CD8 T cells. *Nature* **576**, 465–470 (2019). doi: 10.1038/s41586-019-1836-5; pmid: 31827286
56. L. Gattinoni *et al*., A human memory T cell subset with stem cell-like properties. *Nat. Med.* **17**, 1290–1297 (2011). doi: 10.1038/nm.2446; pmid: 21926977
57. S. K. Vodnala *et al*., T cell stemness and dysfunction in tumors are triggered by a common mechanism. *Science* **363**, eaau0135 (2019). doi: 10.1126/science.aau0135; pmid: 30923193
58. K. L. Tsai *et al*., A conserved Mediator-CDK8 kinase module association regulates Mediator-RNA polymerase II interaction. *Nat. Struct. Mol. Biol.* **20**, 611–619 (2013). doi: 10.1038/nsmb.2549; pmid: 23563140
59. M. T. Knuesel, K. D. Meyer, C. Bernecky, D. J. Taatjes, The human CDK8 subcomplex is a molecular switch that controls Mediator coactivator function. *Genes Dev.* **23**, 439–451 (2009). doi: 10.1101/gad.1767009; pmid: 19240132
60. X. Wang *et al*., Structural flexibility and functional interaction of Mediator Cdk8 module. *Protein Cell* **4**, 911–920 (2013). doi: 10.1007/s13238-013-3069-y; pmid: 24043446
61. B. Aranda-Orgilles *et al*., MED12 regulates HSC-specific enhancers independently of Mediator kinase activity to control hematopoiesis. *Cell Stem Cell* **19**, 784–799 (2016). doi: 10.1016/j.stem.2016.08.004; pmid: 27570068
62. M. D. Galbraith *et al*., CDK8 kinase activity promotes glycolysis. *Cell Rep.* **21**, 1495–1506 (2017). doi: 10.1016/j.celrep.2017.10.058; pmid: 29117556
63. C. Neumayr *et al*., Differential cofactor dependencies define distinct types of human enhancers. *Nature* **606**, 406–413 (2022). doi: 10.1038/s41586-022-04779-x; pmid: 35650434
64. S. M. Kaech *et al*., Selective expression of the interleukin 7 receptor identifies effector CD8 T cells that give rise to long-lived memory cells. *Nat. Immunol.* **4**, 1191–1198 (2003). doi: 10.1038/ni1009; pmid: 14625547
65. E. J. Wherry *et al*., Lineage relationship and protective immunity of memory CD8 T cell subsets. *Nat. Immunol.* **4**, 225–234 (2003). doi: 10.1038/ni889; pmid: 12563257
66. S. C. Wei *et al*., Distinct cellular mechanisms underlie anti-CTLA-4 and anti-PD-1 checkpoint blockade. *Cell* **170**, 1120–1133.e17 (2017). doi: 10.1016/j.cell.2017.07.024; pmid: 28803728
67. R. Firestein *et al*., CDK8 is a colorectal cancer oncogene that regulates β-catenin activity. *Nature* **455**, 547–551 (2008). doi: 10.1038/nature07179; pmid: 18794900
68. J. D. Gillmore *et al*., CRISPR-Cas9 in vivo gene editing for transthyretin amyloidosis. *N. Engl. J. Med.* **385**, 493–502 (2021). doi: 10.1056/NEJMoa2107454; pmid: 34215024
69. C. Ashmore-Harris, G. O. Fruhwirth, The clinical potential of gene editing as a tool to engineer cell-based therapeutics. *Clin. Transl. Med.* **9**, 15 (2020). doi: 10.1186/s40169-020-0268-z; pmid: 32034584
70. E. H. Yau, T. M. Rana, Next-generation sequencing of genome-wide CRISPR screens. *Methods Mol. Biol.* **1712**, 203–216 (2018). doi: 10.1007/978-1-4939-7514-3_13; pmid: 29224076
71. E. R. Zunder *et al*., Palladium-based mass tag cell barcoding with a doublet-filtering scheme and single-cell deconvolution algorithm. *Nat. Protoc.* **10**, 316–333 (2015). doi: 10.1038/nprot.2015.020; pmid: 25612231
72. R. Finck *et al*., Normalization of mass cytometry data with bead standards. *Cytometry A* **83**, 483–494 (2013). doi: 10.1002/cyto.a.22271; pmid: 23512433
73. B. Daniel, B. L. Balint, Z. S. Nagy, L. Nagy, Mapping the genomic binding sites of the activated retinoid X receptor in murine bone marrow-derived macrophages using chromatin immunoprecipitation sequencing. *Methods Mol. Biol.* **1204**, 15–24 (2014). doi: 10.1007/978-1-4939-1346-6_2; pmid: 25182757
74. L. J. Zhu *et al*., ChIPpeakAnno: A Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* **11**, 237 (2010). doi: 10.1186/1471-2105-11-237; pmid: 20459804

ACKNOWLEDGMENTS

Funding: This work was supported by National Institutes of Health grants U54CA232568-01 (C.L.M.), RM1-HG007735 (H.Y.C.), R35-CA209919 (H.Y.C.), and U01CA260852 (A.T.S.); a Stand Up 2 Cancer–St. Baldrick's–National Cancer Institute Pediatric Cancer Dream Team Translational Research Grant (SU2CAACR-DT1113, C.L.M.); Parker Institute for Cancer Immunotherapy (A.T.S., H.Y.C., and C.L.M.); Virginia and D.K. Ludwig Fund for Cancer Research (C.L.M.); the Burroughs Welcome Fund Career Award for Medical Scientists (A.T.S.); a Pew-Stewart Scholars for Cancer Research Award (A.T.S.); and the Cancer Research Institute Lloyd J. Old STAR Award (A.T.S.). Stand Up 2 Cancer is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research. C.L.M., H.Y.C., and A.T.S. are members of the Parker Institute for Cancer Immunotherapy, which supports the Stanford University Cancer Immunotherapy Program. H.Y.C. is an investigator of the Howard Hughes Medical Institute. E.W.W. is a member of the Parker Institute for Cancer Immunotherapy and was funded by a Parker Institute Bridge Scholar Award. R.G.M. is the Taube Distinguished Scholar for Pediatric Immunotherapy at Stanford University School of Medicine. J.A.B. was supported by a Stanford Graduate Fellowship and a National Science Foundation Graduate Research Fellowship under grant DGE-1656518. K.A.F. was supported by the National Science Foundation Graduate Research Fellowship under grant DGE-1656518. Sorting was performed on an instrument in the Stanford Shared FACS Facility obtained using NIH S10 Shared Instrument Grant S10RR025518-01. Tetramers were obtained from the NIH Tetramer Facility, which is supported by contract 75N93020D00005. Illustrations were created with BioRender.com. **Author contributions:** Conceptualization: K.A.F., C.L.M., E.S., A.T.S., R.G.M., and E.W.W. Methodology: K.A.F., J.A.B., B.D., K.S., D.K., P.X., E.W.W., R.G.M., M.M., and A.T.S. Investigation: K.A.F., J.A.B., B.D., K.S., D.K., P.X., M.M., V.T.D., E.S., K.A.B., M.G.D., M.C.R., J.B., P.J.Q., K.A.B., and M.D. Visualization: K.A.F., J.A.B., V.T.D., and V.T. Writing – original draft: K.A.F., E.S., and C.L.M. Writing – review & editing: K.A.F., E.S., C.L.M., E.W.W., H.Y.C., R.G.M., A.T.S., J.A.B., and J.M.E. **Competing interests:** K.A.F., E.S., and C.L.M. are coinventors on patent application number PCT/US2021/058047 submitted by the board of trustees of the Leland Stanford Junior University that covers the use of T cells deficient in *MED12* or *CCNC* for cancer immunotherapy. C.L.M. holds multiple patents in the arena of CAR T cell therapeutics. C.L.M. is a cofounder and holds equity in Lyell Immunopharma, Syncopation Life Sciences, and Link Cell Therapies, which are developing CAR-based therapies, and consults for Lyell, Syncopation, Link, NeoImmuneTech, Apricity, Nektar, Immatics, Mammoth, and Ensoma. A.T.S. is a cofounder of Immunai and Cartography Biosciences. A.T.S. receives research funding from Allogene Therapeutics and Merck Research Laboratories. H.Y.C. is an inventor on patents for the use of ATAC-seq, a cofounder of Accent Therapeutics and Boundless Bio, and an adviser for 10x Genomics, Arsenal Biosciences, Cartography Biosciences, and Spring Discovery. E.W.W. consults for and holds equity in Lyell Immunopharma and VISTAN Health and consults for Umoja Biopharma. E.S. consults for and holds equity in Lyell Immunopharma and consults for Lepton Pharmaceuticals and Galaria. R.G.M. is a cofounder of and holds equity in Syncopation Life Sciences and Link Cell Therapies. R.G.M. has served as a consultant for Lyell Immunopharma, Zai Lab, NKarta, Arovella Pharmaceuticals, Innervate Radiopharmaceuticals, GammaDelta Therapeutics, Immunai, and Aptorum Group. J.A.B. is a consultant to Immunai. **Data and materials availability:** ATAC-seq, ChIP-seq, and RNA-seq data have been deposited into the Gene Expression Omnibus (GEO) repository under accession number GSE174282. ATAC-seq and ChIP-seq data can be viewed through the UCSC Genome Browser at the following link: https://genome.ucsc.edu/s/kfreitas/med12%2Dcart%2Dv3. **License information:** Copyright © 2022 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abn5647
Figs. S1 to S13
Tables S1 to S6
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

Submitted 7 December 2021; resubmitted 2 August 2022
Accepted 22 September 2022
10.1126/science.abn5647

RESEARCH ARTICLE

CORONAVIRUS

Imprinted antibody responses against SARS-CoV-2 Omicron sublineages

Young-Jun Park[1,2]†, Dora Pinto[3]†, Alexandra C. Walls[1,2]†, Zhuoming Liu[4]†, Anna De Marco[3], Fabio Benigni[3], Fabrizia Zatta[3], Chiara Silacci-Fregni[3], Jessica Bassi[3], Kaitlin R. Sprouse[1], Amin Addetia[1], John E. Bowen[1], Cameron Stewart[1], Martina Giurdanella[3], Christian Saliba[3], Barbara Guarino[3], Michael A. Schmid[3], Nicholas M. Franko[5], Jennifer K. Logue[5], Ha V. Dang[6], Kevin Hauser[6], Julia di Iulio[6], William Rivera[6], Gretja Schnell[6], Anushka Rajesh[6], Jiayi Zhou[6], Nisar Farhat[6], Hannah Kaiser[6], Martin Montiel-Ruiz[6], Julia Noack[6], Florian A. Lempp[6], Javier Janer[4], Rana Abdelnabi[7], Piet Maes[7], Paolo Ferrari[9,10,11], Alessandro Ceschi[9,12,13,14], Olivier Giannini[9,15], Guilherme Dias de Melo[16], Lauriane Kergoat[16], Hervé Bourhy[16], Johan Neyts[7], Leah Soriaga[6], Lisa A. Purcell[6], Gyorgy Snell[6], Sean P.J. Whelan[4], Antonio Lanzavecchia[3], Herbert W. Virgin[6,17,18], Luca Piccoli[3], Helen Y. Chu[5], Matteo Samuele Pizzuto[3], Davide Corti[3]*, David Veesler[1,2]*

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) Omicron sublineages carry distinct spike mutations resulting in escape from antibodies induced by previous infection or vaccination. We show that hybrid immunity or vaccine boosters elicit plasma-neutralizing antibodies against Omicron BA.1, BA.2, BA.2.12.1, and BA.4/5, and that breakthrough infections, but not vaccination alone, induce neutralizing antibodies in the nasal mucosa. Consistent with immunological imprinting, most antibodies derived from memory B cells or plasma cells of Omicron breakthrough cases cross-react with the Wuhan-Hu-1, BA.1, BA.2, and BA.4/5 receptor-binding domains, whereas Omicron primary infections elicit B cells of narrow specificity up to 6 months after infection. Although most clinical antibodies have reduced neutralization of Omicron, we identified an ultrapotent pan-variant–neutralizing antibody that is a strong candidate for clinical development.

The emergence of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) Omicron variant at the end of 2021 caused a worldwide surge in COVID-19 cases. The Omicron BA.1 and BA.1.1 lineages swept the world first, followed by the BA.2 lineage (*1*). Although BA.1 and BA.2 share a large number of spike (S) mutations, they are each characterized by unique sets of amino acid changes that are associated with different antigenic properties (*2*–*4*). The BA.2.12.1 sublineage emerged in the United States, peaking at the beginning of June 2022, and is characterized by the presence of the L452Q receptor-binding domain (RBD) and S704L fusion machinery mutations in addition to the BA.2-defining mutations (*4*). The BA.2.75 sublineage is spreading in multiple countries and carries unique mutations (added to the BA.2 background) in the N-terminal domain (NTD), along with D339H, G446S, and N460K mutations and an R493Q reversion in the RBD (*5*). The BA.3 S glycoprotein comprises a combination of mutations found in BA.1 S and BA.2 S (*6*), whereas BA.4 S and BA.5 S are identical to each other and comprise a deletion of residues 69 to 70, L452R and F486V substitutions, and an R493Q reversion compared with BA.2 S (*7*). We characterized the emergence of Omicron (BA.1) as a major antigenic shift because of the unprecedented magnitude of immune evasion associated with this variant of concern (*3*, *8*–*12*). Mutations in the BA.1 S glycoprotein NTD and RBD, which are the main targets of neutralizing antibodies (*3*, *8*, *13*–*18*), explain the markedly reduced plasma-neutralizing activity of previously infected or vaccinated subjects (especially those who have not received booster doses) and the escape from most monoclonal antibodies (mAbs) used in the clinic. As a result, an increasing number of reinfections or breakthrough infections are occurring (*19*–*22*), even though these cases tend to be milder than infections of immunologically naive individuals.

Characterization of plasma and mucosal humoral responses to Omicron infection

Understanding the relationships between prior antigen exposure through vaccination or infection with one SARS-CoV-2 strain and the immune response to subsequent infections with a different strain is paramount to guiding strategies to end the COVID-19 pandemic. To investigate this, we first evaluated the magnitude of immune evasion associated with the Omicron sublineages by assessing the neutralizing activity of human plasma using a nonreplicative vesicular stomatitis virus (VSV) pseudotyped with Wuhan-Hu-1 S harboring G614 (Wu-G614), Delta, BA.1, BA.2, BA.2.12.1, or BA.4/5 mutations or with SARS-CoV S (Fig. 1A; fig. S1, A to G; table S1; and data S1). We compared plasma from six cohorts of individuals: those previously infected in 2020 (with a Washington-1–like SARS-CoV-2 strain) and then vaccinated twice ("infected-vaccinated 2 doses") or three times ("infected-vaccinated 3 doses); those who were vaccinated and then experienced either a Delta or an Omicron BA.1 breakthrough infection ("Delta breakthrough 3 doses," "BA.1 breakthrough 2 doses," and "BA.1 breakthrough 3 doses"); and those who had only been vaccinated and boosted ("vaccinated-only 3 doses"). Neutralizing antibody responses were slightly more robust against BA.2 S VSV than against BA.1 S VSV among all groups except for the BA.1 breakthrough cases. Reductions of geometric mean titers (GMTs) relative to Wu-G614 S VSV ranged from 1.4- to 8.2-fold against BA.1 and from 1.6- to 4-fold against BA.2 (Fig. 1A; fig. S1, A to G; table S1; and data S1), which is consistent with recent findings (*4*). BA.2.12.1 S VSV was associated with further reductions of plasma-neutralizing activity relative to BA.2 S VSV, whereas BA.4/5 S VSV had the greatest impact of all of the SARS-CoV-2 variants evaluated here, with GMT reductions of 5- to 14-fold relative to Wu-G614 S VSV (Fig. 1A; fig. S1, A to G; table S1; and data S1). All six cohorts experienced reductions in plasma-neutralizing GMT of 1.4- to 3.6-fold against Delta (*23*–*25*) relative to Wu-G614 S VSV, underscoring that even hybrid immunity [i.e., that acquired through vaccination and infection (*26*)] does not overcome evasion from neutralizing antibody responses of this previously dominant variant of concern (Fig. 1A; fig. S1, A to G; table S1; and data S1). The highest levels of neutralizing GMTs against SARS-CoV-2 variants were observed for BA.1 breakthrough cases, which was possibly due

[1]Department of Biochemistry, University of Washington, Seattle, WA, USA. [2]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [3]Humabs Biomed SA, Subsidiary of Vir Biotechnology, Bellinzona, Switzerland. [4]Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO, USA. [5]Division of Allergy and Infectious Diseases, University of Washington, Seattle, WA, USA. [6]Vir Biotechnology, San Francisco, CA, USA. [7]KU Leuven Department of Microbiology, Immunology and Transplantation, Rega Institute for Medical Research, Laboratory of Virology and Chemotherapy, B-3000 Leuven, Belgium. [8]Laboratory of Clinical and Epidemiological Virology, Rega Institute, Department of Microbiology, Immunology and Transplantation, KU Leuven, Leuven, Belgium. [9]Faculty of Biomedical Sciences, Università della Svizzera italiana, Lugano, Switzerland. [10]Division of Nephrology, Ente Ospedaliero Cantonale, Lugano, Switzerland. [11]Clinical School, University of New South Wales, Sydney, New South Wales, Australia. [12]Clinical Trial Unit, Ente Ospedaliero Cantonale, Lugano, Switzerland. [13]Division of Clinical Pharmacology and Toxicology, Institute of Pharmacological Sciences of Southern Switzerland, Ente Ospedaliero Cantonale, Lugano, Switzerland. [14]Department of Clinical Pharmacology and Toxicology, University Hospital Zurich, Zurich, Switzerland. [15]Department of Medicine, Ente Ospedaliero Cantonale, Bellinzona, Switzerland. [16]Institut Pasteur, Université Paris Cité, Lyssavirus Epidemiology and Neuropathology Unit, F-75015 Paris, France. [17]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO, USA. [18]Department of Internal Medicine, UT Southwestern Medical Center, Dallas, TX, USA.
***Corresponding author. Email: dveesler@uw.edu (D.V.); dcorti@vir.bio (D.C.)**
†These authors contributed equally to this work.

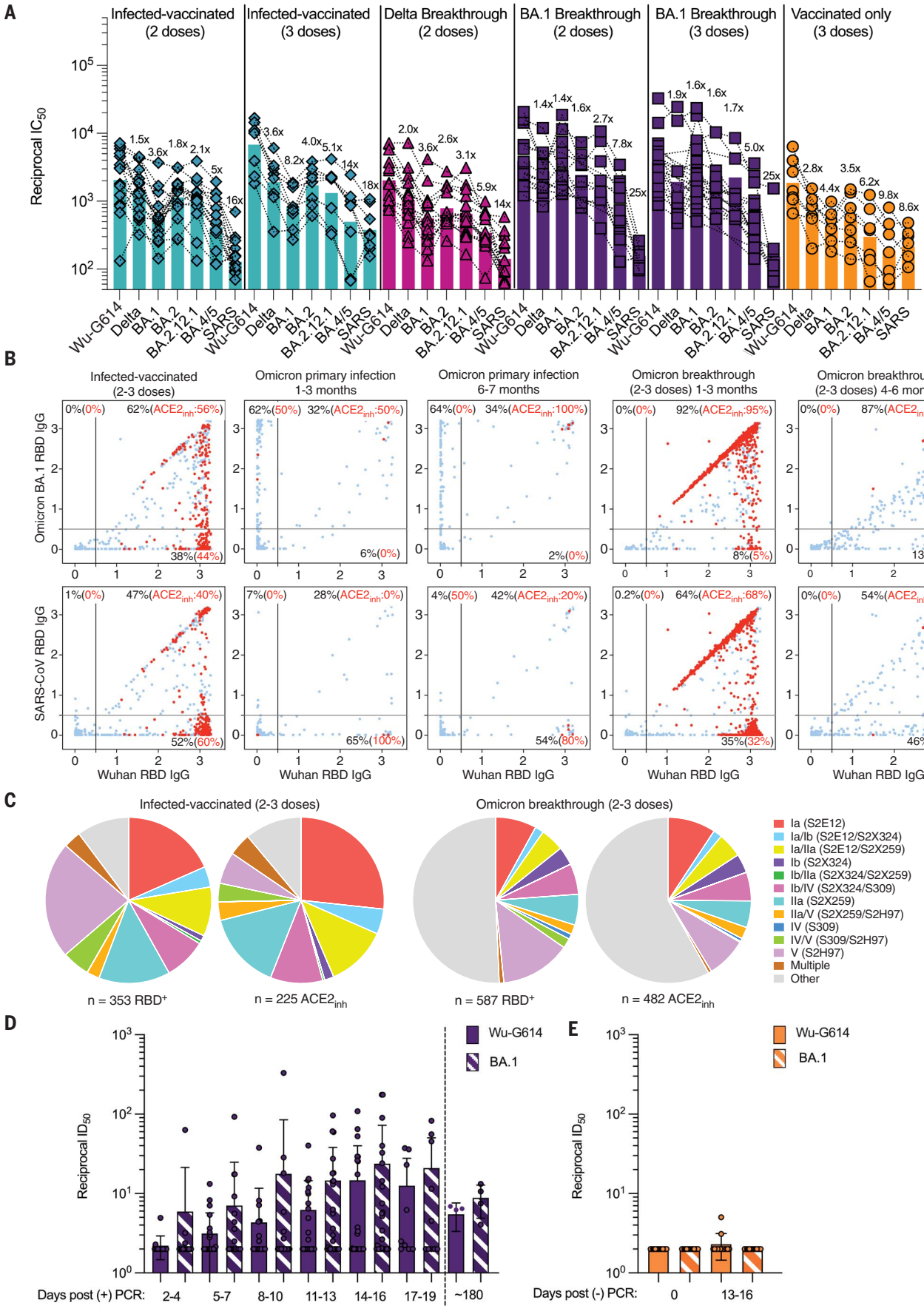

A
Infected-vaccinated (2 doses)
Infected-vaccinated (3 doses)
Delta Breakthrough (2 doses)
BA.1 Breakthrough (2 doses)
BA.1 Breakthrough (3 doses)
Vaccinated only (3 doses)
Reciprocal IC50
Wu-G614
Delta
BA.1
BA.2
BA.2.12.1
BA.4/5
SARS
B
Infected-vaccinated (2-3 doses)
Omicron primary infection 1-3 months
Omicron primary infection 6-7 months
Omicron breakthrough (2-3 doses) 1-3 months
Omicron breakthrough (2-3 doses) 4-6 months
Omicron BA.1 RBD IgG
SARS-CoV RBD IgG
Wuhan RBD IgG
C
Infected-vaccinated (2-3 doses)
Omicron breakthrough (2-3 doses)
n = 353 RBD+
n = 225 ACE2inh
n = 587 RBD+
n = 482 ACE2inh
Ia (S2E12)
Ia/Ib (S2E12/S2X324)
Ia/IIa (S2E12/S2X259)
Ib (S2X324)
Ib/IIa (S2X324/S2X259)
Ib/IV (S2X324/S309)
IIa (S2X259)
IIa/V (S2X259/S2H97)
IV (S309)
IV/V (S309/S2H97)
V (S2H97)
Multiple
Other
D
Reciprocal ID50
Wu-G614
BA.1
Days post (+) PCR:
2-4
5-7
8-10
11-13
14-16
17-19
~180
E
Reciprocal ID50
Wu-G614
BA.1
Days post (-) PCR:
0
13-16

Fig. 1. Evaluation of plasma, memory, and mucosal antibody responses upon Omicron breakthrough infections in humans. (**A**) Pairwise neutralizing activity [half-maximum inhibitory dilution (ID_{50})] against Wu-G614, Delta, BA.1, BA.2, BA.2.12.1, BA.4/5, and SARS-CoV S VSV pseudoviruses using plasma from subjects who were infected and vaccinated, vaccinated and experienced breakthrough infection, or received vaccination only. VeroE6-TMPRSS2 cells were used as target cells (*93*). Data are the geometric mean of an $n = 2$ technical replicates and have been performed in at least two biologically independent experiments. GMTs are shown with a color-matched bar (and reported in table S1) with the fold change compared with Wu-G614 indicated above. Demographics of enrolled donors are provided in data S1. (**B**) Cross-reactivity of IgGs secreted from memory B cells obtained from infected-vaccinated individuals ($n = 11$), primary SARS-CoV-2 infected individuals ($n = 3$ samples collected at 1 to 3 months and $n = 2$ samples collected at 6 to 7 months), or breakthrough cases ($n = 7$ samples collected at 1 to 3 months and $n = 4$ samples collected at 4 to 6 months) occurring in January–March 2022, when the prevalence of Omicron BA.1/BA.2 exceeded 90% in the region where samples were obtained (fig. S2). Each dot represents a well containing oligoclonal B cell supernatant screened for the presence of IgGs binding to the SARS-CoV-2 Wuhan-Hu-1 and BA.1 RBDs (top) or to the SARS-CoV-2 Wuhan-Hu-1 and SARS-CoV RBDs (bottom) using ELISA. Red dots indicate inhibition of the interaction with ACE2 (using Wuhan-Hu-1 target antigen) as determined in a separate assay. The percentages are expressed relative to the total positive hits against any of the antigens tested. Numbers of positive hits relative to individual donors are shown in fig. S3. (**C**) Frequency analysis of site-specific IgG antibodies derived from memory B cells. RBD sites targeted by IgG derived from memory B cells were defined by a blockade-of-binding assay using mAbs specific for sites Ia (S2E12), Ib (S2X324), IIa (S2X259), IV (S309; parent of sotrovimab), and V (S2H97). Hybrid sites Ia/Ib, Ia/IIa, Ib/IIa, Ib/IV, IIa/V, and IV/V were defined by competition with the two corresponding mAbs. Hybrid sites exhibiting competition with more than two mAbs are indicated as "multiple." Lack of competition is indicated as "other." Pie charts show cumulative frequencies of IgGs specific for the different sites among total RBD-directed IgG antibodies (left) and those inhibiting binding of RBD to human ACE2 (right) in $n = 11$ infected-vaccinated individuals or $n = 7$ breakthrough cases. (**D**) Neutralizing activity against Wu-G614 and BA.1 S VSV pseudoviruses determined from nasal swabs obtained longitudinally upon BA.1 breakthrough infection up to approximately 180 days after a positive PCR test [post (+) PCR]. (**E**) Neutralizing activity against Wu-G614 and BA.1 S VSV pseudoviruses from nasal swabs obtained longitudinally after a negative PCR test [post (–) PCR)] in vaccinated-only individuals.

to exposure to BA.1 S because no correlation was found between time intervals and GMTs (data S1). Neutralizing GMTs against the SARS-CoV S pseudovirus was reduced for all cohorts by 8.6- to 25-fold relative to Wu-G614 S VSV, underscoring the marked genetic and antigenic divergence of this sarbecovirus clade (*19*, *27*, *28*).

Given the recall of Wuhan-Hu-1 plasma-neutralizing antibodies in Omicron breakthrough cases, we investigated the cross-reactivity of RBD-directed antibodies produced by in vitro–stimulated memory B cells obtained up to 200 days after infection or vaccination, as well as in circulating plasma cells collected in the days after infection (*29*). These analyses used blood samples from individuals who were infected before the emergence of Omicron and subsequently vaccinated ("infected-vaccinated 2/3 doses"), as well as subjects who experienced either an Omicron primary infection or an Omicron breakthrough infection. Primary and breakthrough Omicron infections occurred between January and March 2022, during which time the prevalence of Omicron BA.1/BA.2 sublineages exceeded 90% in the region from which the samples were obtained (fig. S2). Plasma-neutralizing activity of Omicron-infected (primary and breakthrough) cases was reduced an average of 6.1-fold against BA.4/BA.5 S VSV relative to BA.1 S VSV (table S2), likely as a result of both RBD and NTD mutations in the former lineage, concurring with the above data and recent studies (*30*, *31*). More than 80% of SARS-CoV-2 RBD-directed IgGs secreted by memory B cells and plasma cells obtained from breakthrough cases cross-reacted with the Wuhan-Hu-1, BA.1, BA.2, BA.4/5 and Delta RBDs, and >90% of these antibodies blocked binding to ACE2 [a correlate of neutralization (*13*, *32*)] (Fig. 1B, figs. S3 to S6, and table S2). Moreover, Omicron breakthrough infections failed to elicit BA.1-, BA.2-, or BA.4/5-specific RBD-directed memory B cells. Notably, a fraction of RBD-directed antibodies (7 to 9%) cross-reacted with the Wuhan-Hu-1 and BA.2 RBDs but not with the BA.1 RBD, and a smaller fraction (1 to 3%) also cross-reacted with the Wuhan-Hu-1 and BA.4/5 RBDs but not with the BA.1 RBD, consistent with the antigenic distance of BA.1 from the other Omicron sublineages (Fig. 1B, figs. S3 to S6, and table S2). Furthermore, the proportion of BA.4/5-reacting antibodies cross-reacting with Wuhan-Hu-1, BA.1, and BA.2 decreased over time when comparing 1 to 3 months versus 4 to 6 months after breakthrough infections (fig. S4, D to F). This suggests that the maturation of antibodies driven by BA.1 or BA.2 breakthrough infections may also result in a narrowing of their specificity over time, thereby decreasing cross-reactivity with the BA.4/5 RBD. These findings illustrate how immunological imprinting from prior exposure, also referred to as "original antigenic sin," can strongly affect the response to distantly related antigens. By contrast, memory B cell–derived RBD-directed IgG antibodies obtained from Omicron primary infections up to 6 to 7 months after infection were present at low frequency and were mostly specific for the BA.1 and BA.2 RBDs, (Fig. 1B, figs. S3 to S6, and data S1). The frequency of IgG antibodies cross-reacting with the SARS-CoV RBD was similar across all three cohorts, concurring with the overall weak plasma-neutralizing activity (Fig. 1, A and B, and table S2).

We determined the site specificity of RBD-directed antibodies secreted by stimulated memory B cells by competition with structurally characterized mAbs targeting four distinct antigenic sites (*13*, *27*). Most of the memory B cell–derived antibodies from (pre-Omicron) infected-vaccinated individuals competed with the five reference mAbs used, whereas a large fraction of antibodies from Omicron breakthrough cases did not compete with any of these five mAbs, indicating that they recognize other undefined RBD antigenic sites (Fig. 1C and fig. S7). Antibodies recognizing most antigenic sites overlapping with the receptor-binding motif (RBM), such as mAb S2E12 (*33*), were found at lower frequency upon Omicron breakthrough infections relative to infected-vaccinated subjects, consistent with the presence of several immune escape mutations in the Omicron RBM (Fig. 1C and fig. S7) (*3*, *18*). A similar relative reduction was observed for antibodies targeting RBD antigenic site IIa [recognized by the S2X259 mAb (*34*)] (Fig. 1C and fig. S7), in agreement with previous findings describing Omicron immune escape from several site IIa mAbs (*3*, *8*, *18*). Collectively, these findings demonstrate that Omicron breakthrough infections preferentially expand existing B cell pools primed by vaccination and elicit cross-reactive antibodies, supporting the concept of immunological imprinting.

To evaluate mucosal antibody responses in subjects who experienced a BA.1 breakthrough infection or in vaccinated-only subjects, we assessed IgG- and IgA-binding titers in nasal swabs obtained longitudinally after polymerase chain reaction (PCR) testing. Although we detected S-specific IgG, and to a lesser extent IgA, in swabs from several breakthrough cases, vaccinated-only individuals had no detectable binding antibody titers (fig. S8, A to D, and fig. S9, A and B). Accordingly, we observed mucosal neutralizing activity against Wu-G614 and BA.1 S VSV pseudoviruses for nasal swabs obtained from breakthrough cases throughout the month after symptom onset, corresponding to up to 19 days after positive PCR testing (Fig. 1, D to E; fig. S9C; and data S1). Furthermore, analysis of nasal swabs obtained from four breakthrough cases ~6 months after symptom onset demonstrated a retention of neutralizing activity.

Assessing plasma-neutralizing antibody titers of these BA.1 breakthrough cases yielded similar magnitude and GMT reductions compared with the rest of the BA.1 breakthrough cohort (Fig. 1A, fig. S1F, and data S1). The magnitude of the neutralizing antibody responses in nasal swabs cannot be directly compared with plasma samples because of the self-administration procedure and resulting sample nonuniformity. Overall, we observed heterogeneous mucosal neutralizing antibody responses among BA.1 breakthrough cases but not in vaccinated-only individuals (Fig. 1, D and E; fig. S9, C and D; and data S1). Collectively, these data underscore the lack of or very weak induction of mucosal antibody responses upon intramuscular delivery of mRNA vaccines or adenovirus-vectored vaccines (*35*, *36*) and are consistent with concurrent findings that Omicron breakthrough infection, but not vaccination alone, induces neutralizing antibody responses and tissue-resident T cells in the nasal mucosa (*37*, *38*).

Omicron sublineages escape neutralization mediated by most clinical mAbs

We next evaluated the impact of BA.1, BA.2, BA.3, BA.4, BA.5, BA.2.12.1, and BA.2.75 S mutations on neutralization mediated by a panel of RBD-directed mAbs using VSV pseudoviruses and VeroE6 target cells. The site Ib COV2-2130 mAb weakly neutralized BA.1 (*3*), whereas it neutralized BA.2, BA.3, BA.4, BA.5, BA.2.12.1, and BA.2.75 S VSV pseudoviruses with 1.6-, 4.2-, 14.5-, 8.8-, 2.0-, and 7.9-fold decreases, respectively, in half-maximal inhibition concentration (IC_{50}) compared with Wu-D614 S VSV (Fig. 2A and fig. S10, A and B). Moreover, the COV2-2196 + COV2-2130 mAb cocktail had 106.4-, 7.6-, 35-, 92.8-, 46.5-, 9.3-, and 9.1-fold decreases in potency against BA.1, BA.2, BA.3, BA.4, BA.5, BA.2.12.1, and BA.2.75, respectively (Fig. 2A and fig. S10, A and B). Because COV2-2196 weakly inhibited Omicron sublineages (except for BA.2.75, for which the reduction in IC_{50} was 17.3-fold), the neutralizing activity of the cocktail was largely mediated by COV2-2130. Within the COV2-2130 epitope, position 446 is a glycine residue for Wuhan-Hu-1, BA.2, BA.4, BA.5, and BA.2.12.1 S or a serine residue in BA.1, BA.3, and BA.2.75 S, the latter residue disrupting the binding interface of COV2-2130 (*18*). The importance of this site was also identified through deep mutational scanning (*39*), and this point mutation was shown to reduce neutralizing activity by ~4-fold for COV2-2130 (*8*). The greater reduction in potency against BA.4 and BA.5 relative to BA.2 is likely driven by the L452R mutation, as reported (https://www.fda.gov/media/154701/download) (*39*). The REGN10987 + REGN10933 and LY-CoV16 + LY-CoV555 mAb cocktails and the CT-P59 and ADI-58125 mAbs had reductions of in vitro neutralization potency ranging between two and four orders of magnitude against all Omicron sublineage S VSV pseudoviruses compared with Wu-D614 S VSV because of mutations in the RBM (Fig. 2A and fig. S10, A and B) (*18*). CT-P59, however, retained neutralizing activity against the BA.2.75 sublineage (29.2-fold reduction relative to Wu-D614 S VSV). The recently described ACE2-mimicking S2K146 mAb (*40*), which retained unaltered activity against BA.1 compared with Wu-D614 (*3*), had a mildly reduced neutralizing activity against BA.2, BA.3, BA.2.12.1, and BA.2.75 S VSV pseudoviruses (3.3-, 3.1-, 1.9-, and 4.3-fold, respectively) (Fig. 2A and fig. S10, A and B). However, S2K146 had a marked reduction in neutralizing activity against BA.4 and BA.5 (with 472- and 285-fold IC_{50} reductions compared with Wu-D614 S VSV), likely caused by the F486V mutation.

Sotrovimab, a site IV mAb with broad sarbecovirus (clade Ia and Ib) cross-neutralizing activity (*41*), had a 16-, 7.3-, 21.3-, 22.6-, 16.6-, and 8.3-fold reduction in potency relative to Wu-D614 against VSV pseudoviruses expressing BA.2, BA.3, BA.4, BA.5, BA.2.12.1, and BA.2.75 S proteins, respectively (Fig. 2A). Similar reductions in neutralizing activity were also observed against authentic Omicron sub-lineage virus isolates (Fig. 2C and fig. S11), and are greater than that observed against BA.1 pseudovirus (2.7-fold), although no additional residue mutations map to the sotrovimab epitope except the G339H substitution present in BA.2.75 instead of G339D found in BA.1 (*41*–*43*). We recently showed that sotrovimab retained in vitro effector functions against BA.2 and conferred Fc-dependent protection in the lungs of mice infected with BA.2 (*44*). The additional loss of neutralization of these Omicron sublineage VSV pseudoviruses beyond BA.1 likely results from the S371F substitution, which is found in BA.2, BA.3, BA.4/5, BA.2.12.1, and BA.2.75, and introduces a bulky phenylalanine near the N343 glycan, which is part of the sotrovimab epitope (*41*). A recently determined BA.2 S structure shows that the RBD helix comprising residues 364 to 372 is indeed remodeled (*45*) and adopts a distinct conformation from the ones observed for Wuhan-Hu-1 S or BA.1 S structures (*18*, *46*). This structural rearrangement is sterically incompatible with the glycan N343 conformation observed in S309-bound spike structures (*18*, *41*), as supported by molecular dynamics simulations, and likely explains the reductions in neutralization potency (fig. S11, A to D). Although we could not test the effect of the S371F substitution alone in the Wu-D614 S background (because of poor VSV pseudovirus infectivity), the S371F, S373P, S375F, and D614G mutant (as found in BA.2, BA.3, BA.4, BA.5, BA.2.12.1, and BA.2.75) reduced sotrovimab-mediated neutralization by 3.4-fold relative to Wu-D614 S VSV (fig. S11E and table S3). Moreover, the S371L, S373P, and S375F triple mutant (as found in BA.1) did not alter sotrovimab activity (fig. S11F and table S3), lending further support to the role of F371 in reducing the sotrovimab potency against BA.2, BA.3, BA.4, BA.5, BA.2.12.1, and BA.2.75.

S2X259, a site IIa mAb that broadly reacts with the RBD of multiple sarbecoviruses (*34*), retained activity against BA.1 (*3*). However, the neutralization potency of S2X259 was decreased by one to two orders of magnitude against BA.2, BA.3, BA.4, BA.5, BA.2.12.1, and BA.2.75 S VSV pseudoviruses (Fig. 2A and fig. S10, A and B), likely because of the detrimental effect of the aforementioned S371F/S373P/S375F–induced remodeling and of the R408S mutation (*34*). S2H97 is a site V mAb that had a 4.7- to 10-fold decrease in neutralization potency against Omicron sublineages compared with Wu-D614 S VSV (Fig. 2A and fig. S10, A and B) despite the absence of mutations present in the epitope or otherwise found to affect binding by DMS, perhaps reflecting differential accessibility to its cryptic epitope in the context of these S trimers (*27*).

Identification of the pan-variant and ultrapotent neutralizing mAb S2X324

The S2X324 mAb stood out in our panel because its neutralization potency was largely unaffected by the BA.1, BA.2, BA.3, BA.4, BA.5, BA.2.12.1, and BA.2.75 S mutations (Fig. 2A and fig. S10, A and B). S2X324 cross-reacted with and neutralized all SARS-CoV-2 (VSV pseudovirus and authentic virus) variants tested, with IC_{50} values <10 ng/ml except BA.2.75, for which the IC_{50} was 18 ng/ml (Fig. 2, B and C; figs. S10, A to C, S12, and S13; and table S4). S2X324 cross-reacted with the sarbecovirus clade 1b Pangolin-GD RBD but did not recognize more divergent sarbecovirus RBDs (Fig. 2D), in contrast to the previously described broadly neutralizing mAb S2X259 (*34*). Furthermore, S2X324 inhibited binding of the SARS-CoV-2 RBD to human ACE2 in a concentration-dependent manner, as measured by competition enzyme-linked immunosorbent assay (ELISA) (Fig. 2E), and induced slow, premature shedding (*47*) of the S_1 subunit from cell surface–expressed S (Fig. 2F). However, S2X324 did not promote the fusogenic conformational changes of a wild-type–like purified recombinant S ectodomain trimer (fig. S14), likely because of the slow kinetics of S_1 shedding. This suggests that blockage of ACE2 binding is the main mechanism of S2X324-mediated inhibition of SARS-CoV-2.

To evaluate the ability of S2X324 to promote antibody dependent-phagocytosis or cytotoxicity, we tested whether the mAb could activate Fcγ receptors expressed at the surface of Jurkat cells. Although S2X324 only activated FcγRIIIa, but not FcγRIIa, in vitro (fig. S15, A and B), it triggered both antibody-dependent phagocytosis and cytotoxicity after incubation of

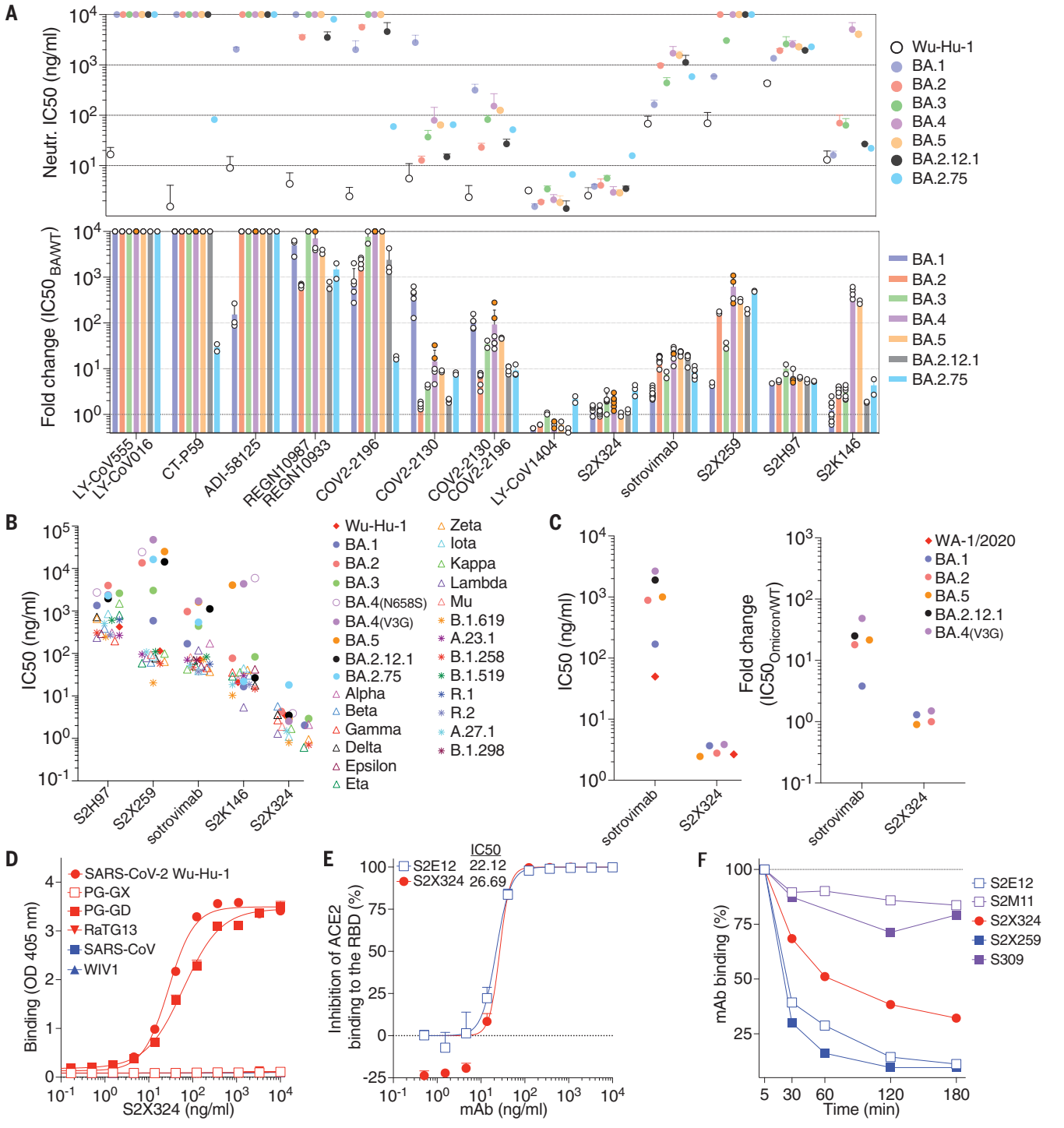


Fig. 2. Identification and characterization of S2X324 as a pan-variant RBD-directed mAb. (**A**) mAb-mediated neutralization of BA.1, BA.2, BA.3, BA.4, BA.5, BA.2.12.1, and BA.2.75 S VSV pseudoviruses. Two haplotypes of BA.4 S were tested: BA.4-V3G (orange dots) and BA.4-N658S (white dots), and the IC_{50} values reported in the text are the averages of both haplotypes. The potency of each mAb or mAb cocktail is represented by their IC_{50} (top, geometric mean ± SD) or fold change relative to neutralization of the Wuhan-Hu-1 (D614) pseudovirus (bottom, average ± SD). (**B**) Neutralization of SARS-CoV-2 variant S VSV pseudoviruses mediated by broadly neutralizing mAbs. Each symbol represents the GMT of at least two independent experiments. (**C**) Neutralizing activity (left) and fold change relative to WA-1/2020 (right) of S2X324 and sotrovimab against SARS-CoV-2 Omicron BA.1, BA.2, BA.4, BA.5, and BA.2.12.1 authentic viruses using VeroE6-TMPRSS2 target cells. Data are representative of at least two biological independent experiments. Neutralization of Omicron BA.1 by sotrovimab refers to previously published data (*3*). (**D**) Cross-reactivity of S2X324 with sarbecovirus clade 1a and 1b RBDs analyzed by ELISA. PG-GX, Pangolin-Guangxi; PG-GD, Pangolin-Guangdong. (**E**) Preincubation of serial dilutions of S2X324 or S2E12 with the SARS-CoV-2 RBD prevents binding to the immobilized human ACE2 ectodomain in ELISA. Error bars indicate SD between replicates. (**F**) S2X324-mediated S_1 shedding from cell surface–expressed SARS-CoV-2 S as determined by flow cytometry. S2E12 and S2X259 were used as positive controls, and S2M11 and S309 were used as negative controls.

peripheral blood mononuclear cells with SARS-CoV-2 S-expressing cells (fig. S15, C to F). The slow S_1 shedding kinetics likely explain the ability of S2X324 to promote Fc-mediated effector functions.

Structural basis for S2X324-mediated neutralization

To understand the pan-variant S2X324 inhibitory activity, we determined a cryo–electron microscopy structure of the Omicron BA.1 S ectodomain trimer bound to the S2X324 Fab fragment at 3.1-Å resolution (Fig. 3A, fig. S16, and table S5). In our structure, the BA.1 S trimer had three Fabs bound to one closed and two open RBDs. We used focused classification and local refinement of the closed RBD-S2X324 Fab complex to obtain a 3.3-Å structure revealing the molecular details of the binding interface.

S2X324 recognizes an RBD epitope partially overlapping with antigenic sites Ib and IV (Fig. 3, A and B), explaining the observed competition with the S2H14 (*13*) and S309 (sotrovimab parent) (*41*) mAbs (fig. S13B). S2X324 uses all six complementary-determining loops to recognize RBD residues T345, N439, K440, L441, S443, K444, V445, S446, G447, N448, Y449, N450, R498, P499, T500, Y501, G502, Q506, and R509 (Fig. 3C). Consistent with the competition assay, S2X324 overlaps with the RBM on the RBD and would sterically hinder receptor engagement (Figs. 2E and 3D).

The structure explains how this mAb accommodates residues that are mutated in Omicron lineages relative to Wuhan-Hu-1: N440K (BA.1/BA.2/BA.3/BA.4/BA.5/BA.2.12.1/BA.2.75), G446S (BA.1/BA.3//BA.2.75), Q498R (BA.1/BA.2/BA.3/BA.4/BA.5/BA.2.12.1/BA.2.75), and N501Y (BA.1/BA.2/BA.3/BA.4/BA.5/BA.2.12.1/BA.2.75). Specifically, K440 forms a salt bridge with the VL E53 side chain, S446 forms van der Waals interactions with VH R60 and VL S96/S97, whereas R498 forms electrostatic interactions with the VL S96 backbone. Our structure further suggests that the tighter binding of S2X324 to the Wuhan-Hu-1 and BA.2 RBDs relative to BA.1 (fig. S13A) might be caused by G446S, because although the mutation is clearly accommodated, at least one of three favored rotamers for S446 would clash with the Fab. The Y501 backbone forms van der Waals interactions with the VL N32 side chain that are independent of the RBD residue identity at position 501 (explaining retention of neutralization of all Y501-containing variants). S2X324 and LY-CoV1404 share 87 and 91% amino acid sequence identity in their heavy and light chains, respectively, likely explaining their similar binding mode (fig. S17) (*48*), pan-variant neutralizing activity (*49*), and comparable resilience to Omicron sublineage mutations thus far (Fig. 2A).

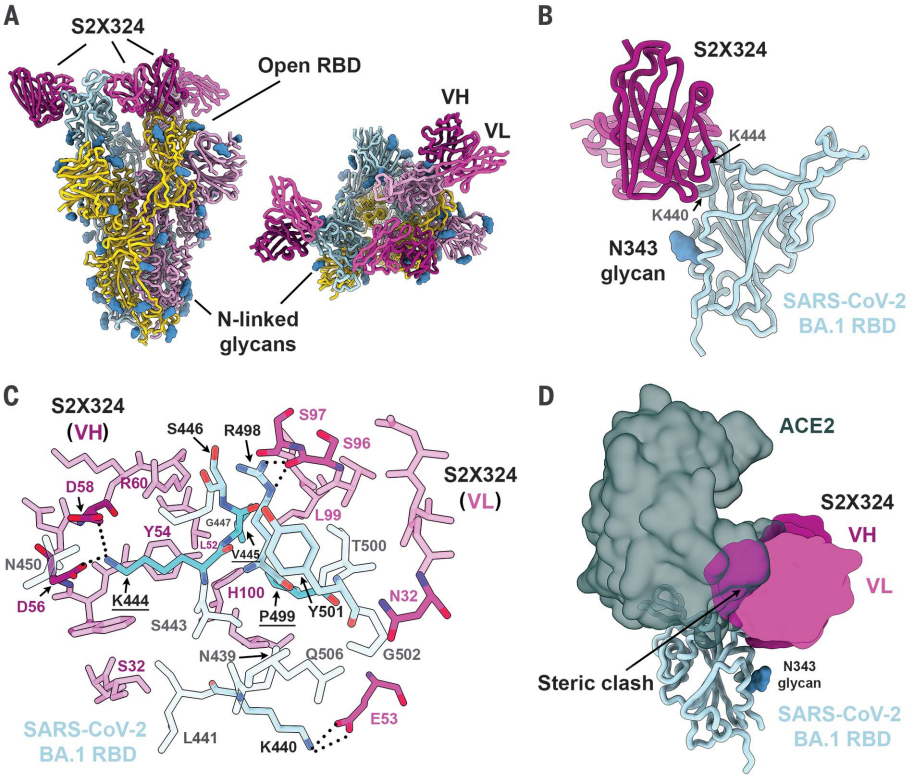


Fig. 3. Structural characterization of the S2X324 pan-variant mAb. (**A**) Cryo-EM structure viewed along two orthogonal orientations of the prefusion SARS-CoV-2 Omicron BA.1 S ectodomain trimer with three S2X324 Fab fragments bound. SARS-CoV-2 S protomers are colored light blue, pink, and gold. S2X324 heavy-chain and light-chain variable domains are colored purple and magenta, respectively. Glycans are shown as blue spheres. (**B**) Ribbon diagram of the S2X324-bound SARS-CoV-2 RBD. The N343 glycan is shown as blue spheres. (**C**) Magnified view of the contacts formed between S2X324 and the SARS-CoV-2 BA.1 RBD. Selected epitope residues are labeled, and electrostatic interactions are indicated with dotted lines. A few of the escape mutants identified are colored turquoise. (**D**) Superimposition of the S2X324-bound (purple and magenta) and ACE2-bound [dark gray, PDB 6M0J (*94*)] SARS-CoV-2 RBD (light blue) structures showing steric overlap. The N343 glycan is shown as blue spheres.

Identification of S2X324 viral escape mutants in vitro

To explore potential mutations that could promote escape from S2X324-mediated neutralization, we passaged a replication-competent VSV chimera harboring either SARS-CoV-2 Wu-G614 S (*50*) or Omicron BA.1 S in the presence of S2X324. Residue substitutions at three distinct sites emerged in both S backgrounds (Fig. 3C; fig. S18, A and B; and tables S6 and S7): (i) K444N/T (Wu-G614 and BA.1 background) and K444E/M (BA.1 background), which would abrogate the salt bridges formed between the K444 side chain and the heavy chain D56 and D58 side chains; (ii) V445D (Wu-G614 background) and V445A/F (BA.1 background), which would disrupt Van der Waals contacts with S2X324; and (iii) P499R (Wu-G614 background) and P499S/H (BA.1 background), which might alter the local RBD backbone conformation and/or sterically hinder mAb binding. Furthermore, three additional mutations were detected in the BA.1 S background only, S446I, G447S, and N448K, which are positioned near the interface between the heavy and light chains (Fig. 3C; fig. S18, A and B; and tables S6 and S7). The VSV chimera harboring SARS-CoV-2 Wu-G614 S outcompeted the chimeras harboring the K444T/N, V445D, or P499R escape mutants after four rounds of passaging, suggesting reduced fitness in this replicating chimeric virus model system (fig. S18C). Even though each of these mutations requires a single nucleotide substitution, they are very rare and have been detected cumulatively only in 0.087 and 0.080% of Delta and Omicron genome sequences as of 12 August 2022, respectively (table S8 and fig. S19), although the frequency of some of them is increasing. We further tested VSV pseudoviruses bearing Wu-G614, BA.1, or BA.2 S carrying K444E, K444D, K444N, K444T, V445D, and P449R/H, and confirmed that these mutations abrogated or strongly reduced S2X324-neutralizing activity (fig. S19 and table S9). In addition, S2X324-neutralizing activity was abrogated when V445T/A/F was introduced in the BA.1 backbone (table S9). S2X324 retained potent neutralizing activity

against pseudoviruses bearing other mutations in the epitope found in known variants such as N439K, N440K, and N501Y in the Wu-G614 S background (table S9). Although the S2X324 escape mutants identified are rare, these data suggest that a mAb cocktail comprising S2X324 would increase the barrier for the emergence of resistance mutants even further compared with this single mAb.

S2X324 protects hamsters against SARS-CoV-2 Delta, BA.2, and BA.5 variants

We investigated the in vivo prophylactic and therapeutic efficacy of S2X324 using Syrian hamsters challenged with SARS-CoV-2 variants. Prophylactic administration of S2X324 or S309 comparably protected hamsters challenged with SARS-CoV-2 Delta in a dose-dependent manner (Fig. 4, A to C) despite a 20-fold difference in in vitro potency against SARS-CoV-2 Delta S VSV (Fig. 2B). These data support the lack of direct correlation between in vitro and in vivo potency that was previously reported (*51*, *52*). Moreover, prophylactic administration of S2X324 at 5 mg/kg decreased viral loads below detection levels in the lungs of hamsters challenged with BA.2 or BA.5 (Fig. 4, D to F). In this model, S309 retained activity against BA.5 despite a 22.6-fold reduced in vitro potency relative to Wu-D614 (Fig. 2, A and B). Therapeutic administration of hamster IgG2a S2X324 (1 day after challenge with the SARS-CoV-2 Delta variant) at 2 and 5 mg/kg prevented body weight loss and reduced lung viral RNA loads by 2.5 and 3 orders of magnitude compared with the control group, respectively (Fig. 4, G and H). Viral replication in the lungs was fully abrogated at 2 and 5 mg/kg of S2X324 and reduced by about one order of magnitude for animals treated with 0.1 and 0.5 mg/kg of S2X324 (Fig. 4I). No statistically significant differences were observed for animals receiving an Fc-silenced version of S2X324 (N297A) versus the groups receiving the same doses of Fc-competent S2X324, indicating that limited contribution of Fc-mediated effector functions in these experimental conditions.

Discussion

Immune imprinting, which is also referred to as original antigenic sin, was described based

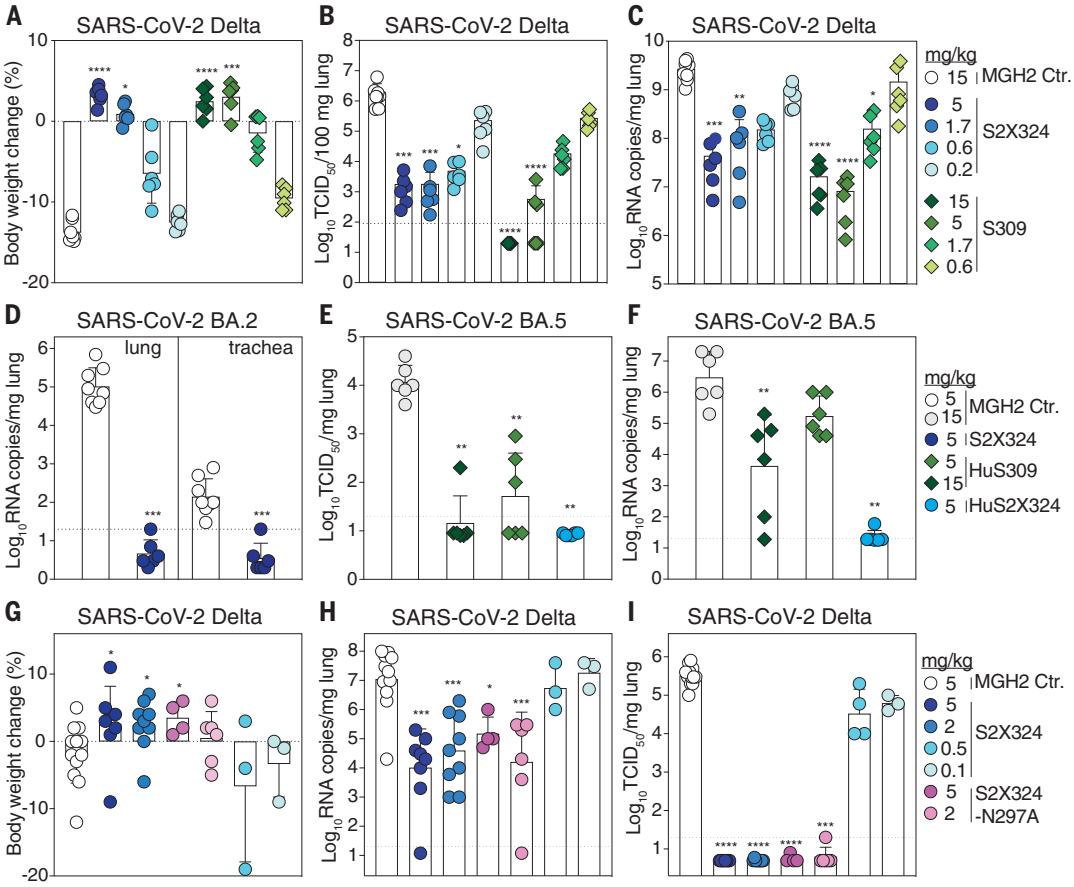


Fig. 4. S2X324 protects hamsters against SARS-CoV-2 Delta, BA.2, and BA.5 challenge. (**A** to **C**) Dose-dependent (expressed in milligrams of mAb per kilogram of body weight) prophylactic protection of S2X324 (blue circles) and S309 (green diamonds) hamster IgG2a (harboring hamster IgG2a constant regions) administered to animals 1 day before infection with SARS-CoV-2 Delta. Animals were evaluated 4 days after infection on the basis of the fraction of body weight change (A), replicating viral titers [50% tissue culture infectious dose ($TCID_{50}$)] (B), and viral RNA load (C). n = 6 animals/dose. *P < 0.05, **P < 0.01, ***P < 0.001, and ****P < 0.0001 relative to isotype control (MGH2 mAb against circumsporozoite protein of Plasmodium sporozoites). Data were analyzed with Kruskal-Wallis test followed by Dunn's multiple-comparisons test. (**D**) Quantification of viral RNA loads in the lung and trachea of Syrian hamsters 4 days after intranasal infection with SARS-CoV-2 Omicron BA.2, which was preceded 1 day prior by prophylactic intraperitoneal administration of S2X324 hamster IgG2a at 5 mg/kg of body weight. ***P < 0.001 relative to control. Data were analyzed with Mann-Whitney two-tailed t test. (**E** and **F**) Quantification of replicating virus titers ($TCID_{50}$) (E) and viral RNA load (F) in the lung of Syrian hamsters 4 days after intranasal infection with SARS-CoV-2 Omicron BA.5, which was preceded 1 day prior by prophylactic intraperitoneal administration of S309 or S2X324 human IgG1 (HuS309 and HuS2X324). (**G** to **I**) Dose-dependent protection in animals 4 days after infection with SARS-CoV-2 Delta by therapeutic intraperitoneal administration of S2X324 hamster IgG2a (blue symbols) or the S2X324 N297A mutant IgG2a (purple symbols) 1 day later at 5, 2, 0.5, or 0.1 mg/kg of body weight. *P < 0.05, **P < 0.01, ***P < 0.001, and ****P < 0.0001 relative to control, respectively. Data were analyzed with Mann-Whitney two-tailed t test.

on the observation that infections with influenza virus strains distinct from the one that caused prior infection preferentially boosted antibody responses against epitopes shared with the original strain (*53*). Although this phenomenon is often considered detrimental, it can also be beneficial, as was the case at the time of the 2009 H1N1 pandemic, during which initial antibody responses to infection with this newly emerged and antigenically shifted virus were dominated by antibodies targeting the conserved hemagglutinin stem region (*54*, *55*). Subsequent exposures through vaccination or infection elicited antibody responses to the shifted variant (i.e., to "non-conserved" hemagglutinin epitopes) (*54*, *56*). Moreover, several studies reported hemagglutinin stem-directed antibody-mediated protection against H5N1 and H7N9 zoonotic influenza strains through imprinting during childhood resulting from exposure to seasonal H1N1 and H3N2, respectively (*55*, *57*). Similarly, we show that exposure to antigenically shifted Omicron strains primarily recalls existing memory B cells specific for epitopes shared by multiple SARS-CoV-2 variants rather than priming naïve B cells recognizing Omicron-specific epitopes (at least up to 180 days after breakthrough infection), as was also recently reported (*58*). Although immune imprinting may be beneficial for stimulating responses to cross-reactive SARS-CoV-2 S epitopes, antibody responses to some Omicron S-specific epitopes were hindered by prior antigenic exposure.

Currently, there is uncertainty whether vaccines matching dominant circulating SARS-CoV-2 variants such as those used for seasonal influenza are needed, or if the repeated use of Wuhan-Hu-1–based vaccines will suffice. Recent work showed that boosting previously immunized macaques with Beta or Omicron mRNA S vaccines or with Beta RBD nanoparticle vaccines elicited comparably high titers of antibodies broadly neutralizing multiple variants relative to Wuhan-Hu-1–based vaccines (*59*–*61*). Furthermore, administration of Wuhan-Hu-1–based vaccine boosters in humans was shown to elicit appreciable titers of neutralizing antibodies and prevent severe disease associated with Omicron infections (*11*, *19*, *62*–*65*). The limited cross-variant neutralization elicited by Omicron primary infection in humans or Omicron-based vaccination of immunologically naïve animals and the data on the specificity of memory B cells presented here indicate that an Omicron-based vaccine might elicit antibody responses directed toward the vaccine-matched and closely related antigens. This suggests that a heterologous prime boost or a multivalent approach might be preferable (*59*, *66*–*73*). Omicron infection and Omicron S-based vaccination of previously immune subjects, however, recalls cross-reactive memory B cells (*58*, *74*), which may further mature over time to enhance their affinity and neutralizing potency against Omicron, but also to possibly broaden their neutralizing activity against past and future variants. Indeed, multiple studies have shown that somatic hypermutations yield RBD-specific mAbs with increased affinity for the homotypic antigen and augmented resilience to immune evasion of emerging heterotypic variants (*40*, *75*–*79*). The recently introduced bivalent mRNA vaccine boosters encoding the Wuhan-Hu-1 and either the BA.1 or the BA.4/5 S glycoproteins have yielded encouraging results (*80*–*82*).

Understanding antibody responses elicited by and directed toward Omicron sublineages is as the result of key to informing public health policies and the design of SARS-CoV-2 and sarbecovirus vaccines (*70*, *71*, *83*–*85*). Our data show that Omicron breakthrough infections do not elicit high titers of pan-sarbecovirus–neutralizing antibodies (e.g., directed against SARS-CoV), in agreement with recent data (*86*). These findings contrast with the observation that preexisting immunity to SARS-CoV followed by SARS-CoV-2 vaccination is associated with elicitation of pan-sarbecovirus–neutralizing antibodies (*28*). These different outcomes might be explained by the low frequency of memory B cells encoding neutralizing antibodies targeting antigenic sites shared among pre-Omicron variants (Wuhan-Hu-1–related strains), Omicron, and SARS-CoV because of the genetic and antigenic distances between these three distinct viruses. For instance, Omicron BA.1 and BA.2 harbor variations of the RBD antigenic site II, which is the target of pan-sarbecovirus–neutralizing antibodies such as S2X259 (*34*), DH1047 (*87*), and ADG2 (*88*), leading to resistance to the neutralization mediated by some of these mAbs (*3*, *8*, *18*). This suggests that conservation of RBD antigenic sites across sarbecoviruses may have resulted (at least partially) from limited immune pressure rather than from functional or structural constraints (i.e., some mutations at these conserved sites may remain compatible with viral fitness) (*86*).

Recent preclinical assessment of intranasally administered influenza and sarbecovirus vaccine candidates has demonstrated the induction of lung-resident protective mucosal humoral and cellular immunity at the site of viral entry (*89*–*92*). These observations, along with our findings that SARS-CoV-2 breakthrough infections, but not vaccination alone, elicit neutralizing activity in the nasal mucosa, support the development and evaluation of a next generation of vaccines administered intranasally.

REFERENCES AND NOTES

1. R. Viana *et al*., *Nature* **603**, 679–686 (2022).
2. J. Yu *et al*., medRxiv 2022.02.06.22270533 [Preprint] (2022); https://doi.org/10.1101/2022.02.06.22270533.
3. E. Cameroni *et al*., *Nature* **602**, 664–670 (2022).
4. J. E. Bowen *et al*., *Science* **377**, 890–894 (2022).
5. C.-W. Tan *et al*., *Lancet Microbe* S2666-5247(22)00220-8 (2022).
6. P. A. Desingu, K. Nagarajan, K. Dhama, *J. Med. Virol.* **94**, 1808–1810 (2022).
7. H. Tegally *et al*., medRxiv 2022.05.01.22274406 [Preprint] (2022); https://doi.org/10.1101/2022.05.01.22274406.
8. L. Liu *et al*., *Nature* **602**, 676–681 (2022).
9. D. Planas *et al*., *Nature* **602**, 671–675 (2022).
10. M. Hoffmann *et al*., bioRxiv 472286 [Preprint] (2021); https://doi.org/10.1101/2021.12.12.472286.
11. W. F. Garcia-Beltran *et al*., *Cell* **185**, 457–466.e4 (2022).
12. H. Gruell *et al*., bioRxiv 487257 [Preprint] (2022); https://doi.org/10.1101/2022.04.06.487257.
13. L. Piccoli *et al*., *Cell* **183**, 1024–1042.e21 (2020).
14. J. E. Bowen *et al*., bioRxiv 473391 [Preprint] (2021); https://doi.org/10.1101/2021.12.19.473391.
15. L. Stamatatos *et al*., *Science* **372**, eabg9175 (2021).
16. A. J. Greaney *et al*., *Sci. Transl. Med.* **13**, eabi9915 (2021).
17. M. McCallum *et al*., *Cell* **184**, 2332–2347.e16 (2021).
18. M. McCallum *et al*., *Science* **375**, 864–868 (2022).
19. A. C. Walls *et al*., *Cell* **185**, 872–880.e3 (2022).
20. A. Y. Collier *et al*., *Sci. Transl. Med.* **14**, eabn6150 (2022).
21. T. A. Bates *et al*., *JAMA* **327**, 179–181 (2022).
22. P. Mlcochova *et al*., Research Square [Preprint] (2021); https://doi.org/10.21203/rs.3.rs-637724/v1.
23. M. McCallum *et al*., *Science* **374**, 1621–1626 (2021).
24. R. Suzuki *et al*., *Nature* **603**, 700–705 (2022).
25. P. Mlcochova *et al*., *Nature* **599**, 114–119 (2021).
26. S. Crotty, *Science* **372**, 1392–1393 (2021).
27. T. N. Starr *et al*., *Nature* **597**, 97–102 (2021).
28. C.-W. Tan *et al*., *N. Engl. J. Med.* **385**, 1401–1406 (2021).
29. D. Pinna, D. Corti, D. Jarrossay, F. Sallusto, A. Lanzavecchia, *Eur. J. Immunol.* **39**, 1260–1270 (2009).
30. K. Khan *et al*., medRxiv 2022.04.29.22274477 [Preprint] (2022).
31. A. Muik *et al*., bioRxiv 502461 [Preprint] (2022); https://doi.org/10.1101/2022.08.02.502461.
32. C. W. Tan *et al*., *Nat. Biotechnol.* **38**, 1073–1078 (2020).
33. M. A. Tortorici *et al*., *Science* **370**, 950–957 (2020).
34. M. A. Tortorici *et al*., *Nature* **597**, 103–108 (2021).
35. L. Azzi *et al*., *EBioMedicine* **75**, 103788 (2022).
36. J. Tang *et al*., *Sci. Immunol.* **7**, eadd4853 (2022).
37. D. Planas *et al*., medRxiv 2022.07.22.22277885 [Preprint] (2022); https://doi.org/10.1101/2022.07.22.22277885.
38. J. M. E. Lim *et al*., *J. Exp. Med.* **219**, e20220780 (2022).
39. J. Dong *et al*., *Nat. Microbiol.* **6**, 1233–1244 (2021).
40. Y.-J. Park *et al*., *Science* **375**, 449–454 (2022).
41. D. Pinto *et al*., *Nature* **583**, 290–295 (2020).
42. E. Cameroni *et al*., bioRxiv 472269 [Preprint] (2021); https://doi.org/10.1101/2021.12.12.472269.
43. A. L. Cathcart *et al*., bioRxiv 434607 [Preprint] (2021); https://doi.org/10.1101/2021.03.09.434607.
44. J. B. Case *et al*., *Nat. Commun.* **13**, 3824 (2022).
45. V. Stalls *et al*., *Cell Rep.* **39**, 111009 (2022).
46. A. C. Walls *et al*., *Cell* **181**, 281–292.e6 (2020).
47. A. C. Walls *et al*., *Cell* **176**, 1026–1039.e15 (2019).
48. K. Westendorf *et al*., bioRxiv 442182 [Preprint] (2022); https://doi.org/10.1101/2021.04.30.442182.
49. K. Westendorf *et al*., *Cell Rep.* **39**, 110812 (2022).
50. J. B. Case *et al*., *Cell Host Microbe* **28**, 475–485.e5 (2020).
51. A. Schäfer *et al*., *J. Exp. Med.* **218**, e20201993 (2021).
52. J. B. Case *et al*., bioRxiv 484787 [Preprint] (2022); https://doi.org/10.1101/2022.03.17.484787.
53. T. Francis, *Proc. Am. Philos. Soc.* **104**, 572–578 (1960).
54. D. Corti *et al*., *Science* **333**, 850–856 (2011).
55. J. Wrammert *et al*., *J. Exp. Med.* **208**, 181–193 (2011).
56. C. S.-F. Cheung *et al*., *Cell Rep.* **32**, 108088 (2020).
57. K. M. Gostic, M. Ambrose, M. Worobey, J. O. Lloyd-Smith, *Science* **354**, 722–726 (2016).
58. J. Quandt *et al*., *Sci. Immunol.* **7**, eabq2427 (2022).
59. M. Gagne *et al*., *Cell* **185**, 1556–1571.e18 (2022).
60. K. S. Corbett *et al*., *Science* **374**, 1343–1353 (2021).
61. P. S. Arunachalam *et al*., *Sci. Transl. Med.* **14**, eabq4130 (2022).
62. E. K. Accorsi *et al*., *JAMA* **327**, 639–651 (2022).
63. H. F. Tseng *et al*., *Nat. Med.* **28**, 1063–1071 (2022).
64. R. Pajon *et al*., *N. Engl. J. Med.* **386**, 1088–1091 (2022).
65. J. E. Bowen *et al*., bioRxiv 484542 [Preprint] (2022); https://doi.org/10.1101/2022.03.15.484542.
66. A. Rössler, L. Knabl, D. von Laer, J. Kimpel, *N. Engl. J. Med.* **386**, 1764–1766 (2022).
67. I.-J. Lee *et al*., bioRxiv 478406 [Preprint] (2022); https://doi.org/10.1101/2022.01.31.478406.
68. S. I. Richardson *et al*., *Cell Host Microbe* **30**, 880–886.e4 (2022).

69. K. Stiasny *et al*., Research Square [Preprint] (2022); https://doi.org/10.21203/rs.3.rs-1536794/v1.
70. A. C. Walls *et al*., *Cell* **184**, 5432–5447.e16 (2021).
71. A. A. Cohen *et al*., *Science* **371**, 735–741 (2021).
72. S. Chalkias *et al*., Research Square [Preprint] (2022); https://doi.org/10.21203/rs.3.rs-1555201/v1.
73. S. S. M. Cheng *et al*., *J. Clin. Virol.* **156**, 105273 (2022).
74. W. B. Alsoussi *et al*., bioRxiv 509040 [Preprint] (2022); https://doi.org/10.1101/2022.09.22.509040.
75. C. Gaebler *et al*., *Nature* **591**, 639–644 (2021).
76. Z. Wang *et al*., *Nature* **595**, 426–431 (2021).
77. D. Pinto *et al*., *Science* **373**, 1109–1116 (2021).
78. J. S. Low *et al*., bioRxiv 486377 [Preprint] (2022); https://doi.org/10.1101/2022.03.30.486377.
79. R. Marzi *et al*., bioRxiv 509852 [Preprint] (2022); https://doi.org/10.1101/2022.09.30.509852.
80. S. Chalkias *et al*., *N. Engl. J. Med.* **387**, 1279–1291 (2022).
81. S. M. Scheaffer *et al*., bioRxiv 507614 [Preprint] (2022); https://doi.org/10.1101/2022.09.12.507614.
82. A. Muik *et al*., bioRxiv 508818 [Preprint] (2022); https://doi.org/10.1101/2022.09.21.508818.
83. A. A. Cohen *et al*., bioRxiv 485875 [Preprint] (2022); https://doi.org/10.1101/2022.03.25.485875.
84. D. R. Martinez *et al*., *Science* **373**, 991–998 (2021).
85. D. Li *et al*., bioRxiv 477915 (2022); https://doi.org/10.1101/2022.01.26.477915.
86. L.-F. Wang *et al*., Research Square [Preprint] (2022); https://doi.org/10.21203/rs.3.rs-1362541/v1.
87. D. R. Martinez *et al*., *Sci. Transl. Med.* **14**, eabj7125 (2022).
88. C. G. Rappazzo *et al*., *Science* **371**, 823–829 (2021).
89. T. Mao *et al*., bioRxiv 477597 [Preprint] (2022); https://doi.org/10.1101/2022.01.24.477597.
90. J. E. Oh *et al*., *Sci. Immunol.* **6**, eabj5129 (2021).
91. S. N. Langel *et al*., *Sci. Transl. Med.* **14**, eabn6868 (2022).
92. A. O. Hassan *et al*., *Cell Rep. Med.* **2**, 100230 (2021).
93. F. A. Lempp *et al*., *Nature* **598**, 342–347 (2021).
94. J. Lan *et al*., *Nature* **581**, 215–220 (2020).

ACKNOWLEDGMENTS

We thank A. E. Powell and N. Czudnochowski for assistance with protein production. **Funding:** This study was supported by the National Institute of Allergy and Infectious Diseases (grants DP1AI158186 and HHSN75N93022C00036 to D.V.), a Pew Biomedical Scholars Award (D.V.), an Investigators in the Pathogenesis of Infectious Disease Award from the Burroughs Wellcome Fund (D.V.), Fast Grants (D.V.), the University of Washington Arnold and Mabel Beckman cryoEM center (D.V.), and the National Institutes of Health (grant S10OD032290 to D.V. and grant AI163019 to S.P.J.W.) D.V. is an Investigator of the Howard Hughes Medical Institute. O.G. is funded by the Swiss Kidney Foundation. **Author contributions:** A.C.W., A.L., D.P., D.C., M.S.P., and D.V. designed the experiments. A.C.W., A.D.M., D.P., C.S., W.R., K.R.S. F.Z., H.V.D., M.G., G.Sc., and F.A.L. isolated mAb and performed binding, neutralization assays, biolayer interferometry, and surface plasmon resonance binding measurements. A.R., J.Z., N.F., M.M.R., and J.N. performed neutralization assays using authentic virus. H.K. confirmed the Spike mutations of authentic virus by Sanger sequencing. A.D.M. and D.P. performed ACE2 binding inhibition and S_1 shedding assays. B.G. and M.A.S. evaluated effector functions. C.S.F., J.B., and L.P performed memory B cell repertoire analysis. O.G., A.C., and P.F. contributed to the recruitment of donors and the collection of plasma samples. J.d.I., L.S., and A.T. performed bioinformatic and epidemiology analyses. Z.L. and S.P.J.W. performed mutant selection and fitness assays. R.A., J.J., F.B., P.M., J.N., G.D.d.M., L.K., and H.B. performed hamster model experiments and data analysis. A.A. performed the EM refolding experiments Y.J.P. prepared cryoEM specimens and collected and processed data collection. Y.J.P. and D.V. built and refined the atomic models. J.E.B. and C.S. purified recombinant glycoproteins. A.L., D.P., Y.J.P., A.D.M., Z.L., D.P., D.C., M.S.P., and D.V. analyzed the data. A.C.W., D.P., D.C., M.S.P., and D.V. wrote the manuscript with input from all authors. F.B., G.S., J.N., S.P.J.W., H.W.V, M.S.P., D.C., and D.V. supervised the project. **Competing interests:** D.P., A.D.M., F.Z., M.G., C.S.F., J.B., C.S., H.V.D., K.H., W.R., M.A.S., G.Sc., B.G., F.B., J.d.I., A.R., J.Z., N.F., H.K., M.M.R, J.N., F.A.L., G.S., L.P., A.T., H.W.V., A.L., M.S.P., and D.C. are employees of Vir Biotechnology Inc. and may hold shares in Vir Biotechnology Inc. L.A.P. is a former employee and shareholder in Regeneron Pharmaceuticals. Regeneron provided no funding for this work. H.W.V. is a founder and holds shares in PierianDx and Casma Therapeutics. Neither company provided resources. D.C. is currently listed as an inventor on multiple patent applications, which disclose the subject matter described in this manuscript. The Veesler laboratory has received a sponsored research agreement from Vir Biotechnology Inc. S.P.J.W. has licensing agreements with Vir Biotechnology and Merck and is a consultant for Thylacine Bio. The remaining authors declare no competing interests. **Data and materials availability:** The cryoEM map and coordinates have been deposited to the Electron Microscopy Databank (SARS-CoV-2 S/S2X324: EMD-28559; SARS-CoV-2 S/S2x324: EMD-28558) and the Protein Data Bank (SARS-CoV-2 S/S2X324: PDB 8ERR; SARS-CoV-2 S/S2x324: PDB 8ERQ). Materials generated in this study will be made available on request, but may require a completed materials transfer agreement signed with Vir Biotechnology Inc. or the University of Washington.

License information: This article is subject to Howard Hughes Medical Institute's (HHMI's) Open Access to Publications policy. HHMI lab heads have previously granted a nonexclusive CC BY 4.0 license to the public and a sublicensable license to HHMI in their research articles. Pursuant to those licenses, the author-accepted manuscript of this article can be made freely available under a CC BY 4.0 license immediately upon publication.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adc9127
Materials and Methods
Figs. S1 to S18
Tables S1 to S9
References (*95*–*126*)
MDAR Reproducibility Checklist
Data S1

View/request a protocol for this paper from *Bio-protocol*.

Submitted 9 May 2022; accepted 17 October 2022
10.1126/science.adc9127

STRUCTURAL BIOLOGY

Structures of a mobile intron retroelement poised to attack its structured DNA target

Kevin Chung[1]†, Ling Xu[2,3]†, Pengxin Chai[1], Junhui Peng[4], Swapnil C. Devarkar[1], Anna Marie Pyle[2,3]*

Group II introns are ribozymes that catalyze their self-excision and function as retroelements that invade DNA. As retrotransposons, group II introns form ribonucleoprotein (RNP) complexes that roam the genome, integrating by reversal of forward splicing. Here we show that retrotransposition is achieved by a tertiary complex between a structurally elaborate ribozyme, its protein mobility factor, and a structured DNA substrate. We solved cryo–electron microscopy structures of an intact group IIC intron-maturase retroelement that was poised for integration into a DNA stem-loop motif. By visualizing the RNP before and after DNA targeting, we show that it is primed for attack and fits perfectly with its DNA target. This study reveals design principles of a prototypical retroelement and reinforces the hypothesis that group II introns are ancient elements of genetic diversification.

Group II introns are self-splicing retroelements that have played a key role in shaping eukaryotic genomes as the ancestors of spliceosomal introns and non–long terminal repeat (LTR) retroelements (*1*). They remain important for gene expression in plants, fungi, yeasts, and many bacteria (*2*, *3*). Group II introns encode a specialized reverse transcriptase (maturase) that binds its parent intron and facilitates self-splicing, which releases a well-folded lariat ribonucleoprotein (RNP) complex (*4*). The liberated RNP functions as a retrotransposon, targeting DNA that contains spliced exon junction sequences and inserting by means of a two-step transesterification reaction known as reverse splicing (Fig. 1A) (*5*). The resulting DNA-RNA chimera is copied into cDNA by the reverse transcriptase (RT) activity of the multifunctional maturase in a process known as target primed reverse transcription (TPRT) (*6*). Host repair pathways complete the downstream DNA copy-and-paste steps that are needed to achieve total intron integration (*4*).

There are three main classes of group II introns, IIA, IIB, and IIC, which share a conserved secondary structure and a similar tertiary organization around a ribozyme active site (*7*). Group IIC introns are an ancient class of bacterial introns that recognize both the sequence and three-dimensional (3D) structure of their DNA insertion sites (*8*). Unlike their larger IIA and IIB counterparts, group IIC introns are almost completely dependent on their maturases to facilitate intron excision through lariat formation, thereby forming the functional RNP that serves as the minimal element for retrotransposition (*8*). Compared with their more evolved counterparts, group IIC RTs lack an endonuclease domain for generating TPRT primers and instead exploit the lagging strands at DNA replication forks (*4*).

Recent structural and biochemical studies of IIA and IIB introns have provided important insights into strategies for the maturase recognition of intron RNA (*9*–*11*). However, available RNP structures have not revealed a specific mechanistic role for the maturase during RNP assembly, DNA recognition, or chemical catalysis. At present, the mechanism by which group IIC introns recognize DNA structures and not just DNA sequences remains unclear. Furthermore, there are no available structures of the free RNP retroelement before it has bound DNA. These open questions preclude a clear understanding of group II intron retrotransposition and its evolutionary role in shaping modern genomes. To address these problems, we solved cryo–electron microscopy (cryo-EM) structures of a group IIC intron retroelement that was poised to undergo the first step of reverse splicing.

Results

Overall architecture of an ancient group II intron retroelement

To investigate the mechanism of DNA insertion, we captured a group II intron retroelement before the first step of reverse splicing into DNA (Fig. 1A). We first conducted in vitro splicing reactions of the IIC *Eubacterium rectale* (*E.r.*) intron (*12*) in complex with its encoded maturase (MarathonRT) (*13*, *14*) and purified the reaction mixture to obtain a branched lariat-maturase complex (fig. S1, A to C). The purity and stability of this RNP complex were assessed by using biophysical methods: sedimentation velocity analytical ultracentrifugation and size exclusion chromatography coupled to multiangle light scattering (SEC-MALS) indicated that the sedimentation coefficient and molecular mass of the RNP were larger compared with those of the individual lariat or maturase components, which suggests complex formation (fig. S1, D to F). To visualize the retroelement in action, we introduced a desthiobiotin-tagged DNA substrate to the intron-maturase RNP and isolated ternary complexes by affinity purification on an avidin column (fig. S2). The purified elution fraction was vitrified on grids, and the holoenzyme molecules appeared as monodisperse particles on cryo-EM micrographs, thereby allowing structure determination (fig. S2, B to D, and fig. S3).

The initial data analysis suggested a preferred orientation of the sample, so a tilted data collection strategy was required to obtain additional projection views (fig. S3). After further classification and focused refinement, we obtained a 2.8-Å resolution cryo-EM structure of the *E.r.* group IIC intron in complex with its specific maturase and DNA target (Fig. 1, B and C, and figs. S4 and S5), thereby revealing the state immediately before the first step of reverse splicing (Fig. 1A). The overall high-resolution 3D reconstruction was of sufficient quality to permit the modeling of individual nucleotides (movie S1) and metal ions. The catalytic core that was formed by D5, the lariat branchpoint, EBS-IBS (exon binding site–intron binding site) sequences, and the protein thumb and DNA binding domain (DBD) was resolved to <3 Å.

The overall structure reveals a compact assembly of intron RNA and maturase protein that is closely associated with the DNA substrate through an extensive network of interactions (movie S2). The intron core adopts a similar fold to that of intron structures that are derived from truncated and modified constructs (*15*, *16*). The tertiary interactions that were identified in previous group II introns are present, along with several additional interactions that are observed in this full-length intron construct that contains all six intron domains (Fig. 1D). The fold of the maturase resembles that of previously studied IIC proteins (*13*, *17*), although the thumb and DNA binding domains are now clearly resolved (Fig. 1C and fig. S5). The bound DNA contains a short spacer, the intron insertion site, and a 5′ stem-loop motif that is exclusive to group IIC introns (Fig. 1D) (*18*, *19*).

Features of the catalytic RNP core

Despite extensive efforts, a complete group II intron holoenzyme active site had not yet been visualized. In this work, we capture the complete ribozyme core architecture, which includes hallmark elements identified in earlier biochemical and structural studies (*8*, *20*). For example, we see that the 2′-5′ lariat linkage, between the first intron nucleotide (G1) and the branchpoint A (A632), is a crucial structural motif for organizing the ribozyme core.

[1]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA. [2]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT 06511, USA. [3]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA. [4]Laboratory of Evolutionary Genetics and Genomics, The Rockefeller University, New York, NY 10065, USA.
*Corresponding author. Email: anna.pyle@yale.edu
†These authors contributed equally to this work.

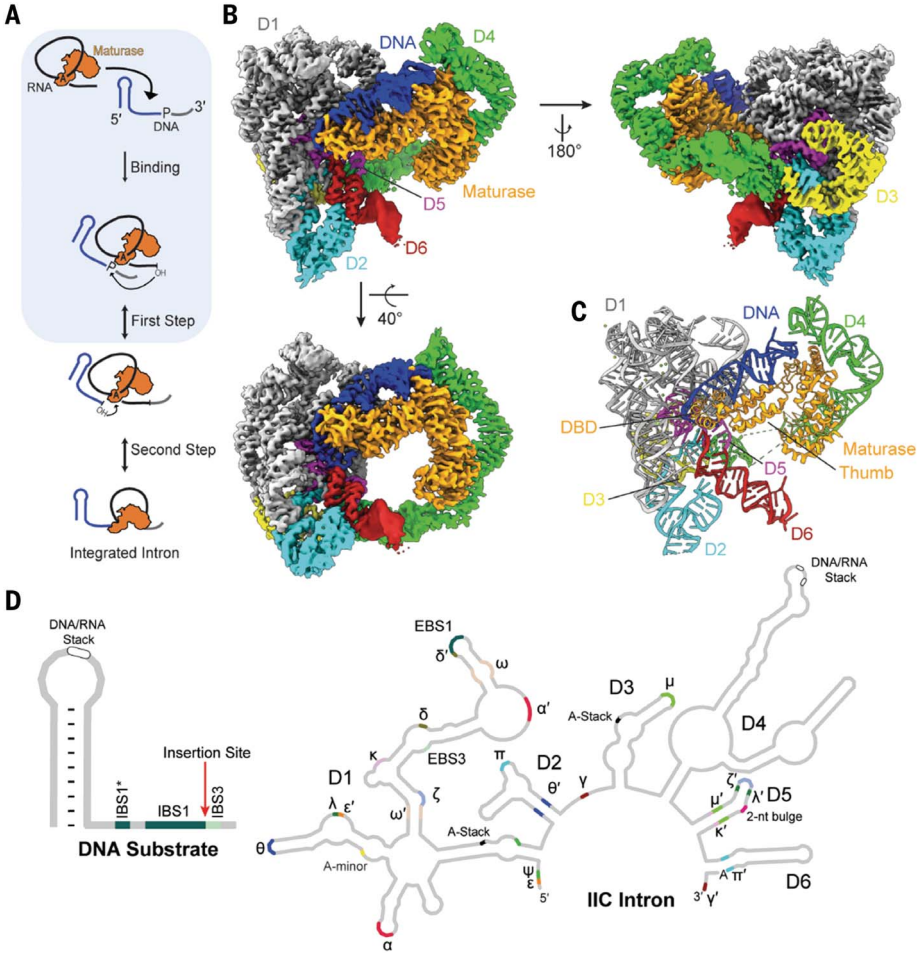


Fig. 1. Cryo-EM reconstruction of a group II intron retroelement. (**A**) Cartoon of the reverse splicing reaction. (**B**) Composite cryo-EM map of the holo-RNP with bound DNA. (**C**) Molecular model of the group II intron retroelement. (**D**) Secondary structure cartoon and tertiary interactions of the holo-RNP.

The branch site actively engages the 3′ end of the intron (G1-A637 pairing), which helps to position the terminal nucleotide (U638) for nucleophilic attack on DNA (Fig. 2, A, B, and E) (*21*, *22*). Facilitating this process, U638 base pairs with A327 to form the γ-γ′ interaction (Fig. 2, E and F) (*21*, *22*). The adjacent G328 and C329 nucleotides of the J2/3 linker form major-groove base triples with C562 and G563 (fig. S6A), which gives rise to the catalytic triplex that is common to all group II introns and the spliceosome (*23*, *24*). The 2-nucleotide (nt) bulge (A580 and C581) and catalytic triad (C562, G563, and C564) in D5, along with U638, all serve to coordinate catalytic magnesium ion M1, placing it between the nucleophilic 3′ OH and scissile phosphate in an arrangement poised for the first step of reverse splicing (Fig. 2, A, B, and E) (*15*). A second magnesium ion, M2, is located 3.9 Å away from M1, which is consistent with the two-metal ion catalysis mechanism (Fig. 2, A, B, and E) (*15*, *25*). We identified two additional, unambiguous densities at positions that were previously assigned to the monovalent ions K1 and K2 in studies that used anomalous scattering to establish sites of stable K^+ binding (*24*, *26*). In that case, as in this instance, NH_4^+ can functionally substitute for K^+ at these same positions (Fig. 2, A, B, and E). The specific coordination and placement of these monovalent ions is essential for positioning the catalytic divalent metal ions, forming a reactive, heteronuclear metal ion cluster. Several tertiary interactions stabilize the periphery of the catalytic core, with D3, supported by an A-stacking interaction with D1, bracing the backside of the D5 helix (μ-μ′) (fig. S6, B to C). D2 contacts D6 (π-π′) to hold the lariat in place (fig. S6D) (*21*, *27*). Although many of these active site elements have been observed independently, in linear introns or in introns of other classes, they have not been captured simultaneously in a single structure until now, thereby demonstrating that these active site elements function in concert and are conserved. The *E.r.* holoenzyme structure provides a detailed view of a complete, reactive intron catalytic core (movie S3).

Close inspection of the active site reveals structural interdependence between the intron and its encoded maturase. Within the active site, the intron RNA forms short base pairings with its target DNA through the EBS-IBS interactions (EBS1-IBS1 and EBS3-IBS3) (Fig. 2, A, B, E, and F). These otherwise unstable short pairing interactions are buttressed and positioned by the maturase, which presses the middle α helices of the DBD and the third α helix of the thumb domain against the EBS1 and EBS3 recognition loops, respectively (Figs. 1B and 2C), which rigidifies them and helps form a central cavity for engagement with DNA (Fig. 2, C and D). These findings establish that the retroelement core does not consist solely of RNA; rather, it is a collaborative, RNP-active site. The previously undescribed roles that we observe for the maturase thumb and DBD help explain the strong maturase dependence for both RNA splicing and intron integration, particularly in vivo, and they highlight the symbiotic relationship between the intron RNA and its protein cofactor, which are known to have coevolved (*28*, *29*).

Functional coordination between RNA and protein

The retroelement holoenzyme has an expansive D4 arm, which extends far from the core and then curves around to cradle the maturase (Fig. 3A). D4a, the high-affinity maturase-binding subdomain (*13*, *30*), forms two anchor points with basic protein surfaces (Fig. 3, A and B) (*31*). At the first anchor point, residues extending from the protein [Arg^{58} (R58), Asp^{152} (D152), Thr^{156} (T156), and R160] interact with RNA phosphate and ribose oxygens to secure the insertion helix within the finger domain (IFD) of the maturase against the minor groove interface in the middle of the long D4a hairpin (Fig. 3, B and D). A sharp turn places the distal portion of the D4a subdomain between α helices 9 and 10 of the protein, where largely basic residues [R217, Ser^{234} (S234), S237, R240, R243, Asn^{244} (N244), and R247] approach the RNA backbone from either side, fastening the palm to the D4a arm (Fig. 3, B and C). In contrast to other group II RNPs, the surface of the finger domain (RT0) is not used for RNA recognition (fig. S7) (*9*, *10*, *13*).

The distinctive intron-maturase recognition strategy places the maturase thumb and DBD next to the intron core, which allows the protein to participate in catalysis by rigidifying the active site (Fig. 3, A and E). The thumb and DBD grasp the EBS1 and EBS3 loops to directly coordinate substrate-recognition elements within the retroelement active site (Fig. 3E). One approach of this strategy involves locking EBS nucleotides into a conformation conducive for substrate binding [i.e., Lys^{388} (K388) with G187O6 of EBS1 and K358 with A231N7 of EBS3] (Fig. 3, E to G). A secondary tactic includes immobilizing the EBS3 phosphate backbone

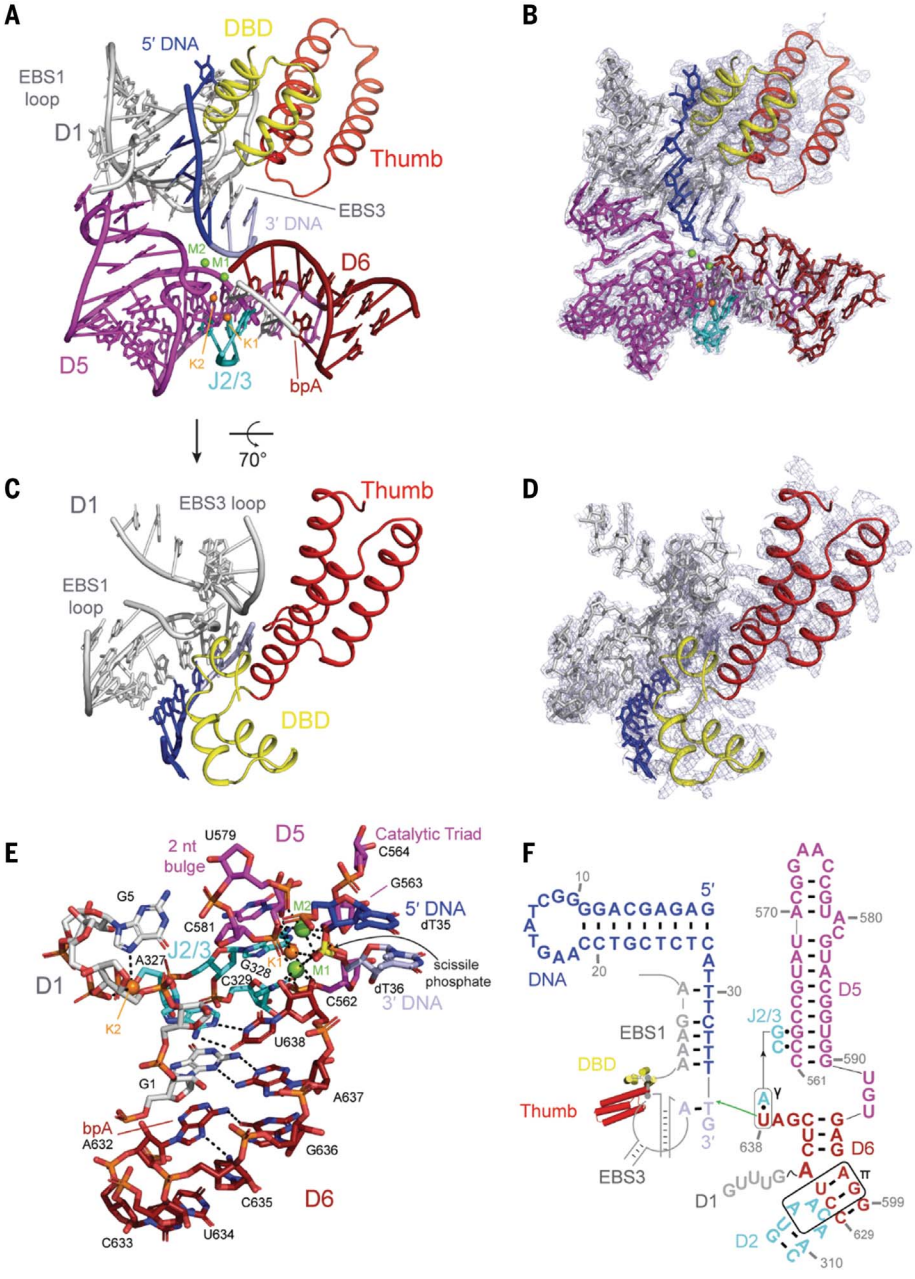


Fig. 2. Architecture of an intron retroelement active site. (**A** and **B**) Organization of the holo-RNP core domains. (**C** and **D**) Maturase DBD and thumb domains stabilize the DNA recognition loops. (**E** and **F**) Model and secondary structure schematic of the intron retroelement before the first step of retrotransposition.

through interactions with a multitude of basic residues on the protein thumb (K300, K303, S309, and R347) (Fig. 3G). A third strategy consists of amino acids [Tyr^{350} (Y350), R389, N395, and main-chain amines of Ile^{390} (I390) and Ala^{391} (A391)] stabilizing the turn in EBS1 and enabling the formation of δ-δ′, thereby reinforcing this single–base pair interaction (C183 with G158) that bridges the EBS loops (Fig. 3F). R308 of the protein thumb provides additional stabilization by simultaneously coordinating the phosphate backbone of EBS1 and -3 (through C183 and A230) (Fig. 3H). These interactions demonstrate a specific mechanistic role for the maturase protein during catalysis, which shows that it promotes proper formation of multiple active site components (*32*). These findings reveal the inextricable, functional coordination of intron and protein during the mechanism of splicing and retrotransposition.

Tertiary interactions with a structured DNA

Our structure reveals unusual strategies for molecular recognition of the DNA target by the holoenzyme. The DNA is recognized through a combination of shape selectivity and base-specific interactions (movie S4), only a few of which involve canonical Watson-Crick (WC) pairing. The DNA itself has distinct structural features that support this recognition strategy. Most prominent is an unusual, structurally conserved DNA stem, which is composed of a short helix [9 base pairs (bp)] that is capped by an undertwisted duplex composed of two non-canonical G-A DNA base pairs and a G-C base pair (Fig. 4, A and B). Together, these extend the DNA stem to 12 bp, which approximates the consensus stem length for IIC insertion targets. The terminal DNA loop serves as a stacking platform for long-range interactions. Adjacent to the DNA stem is a short spacer, which is followed by IBS1 nucleotides and the IBS3 nucleotide that flank the DNA insertion site.

The DNA stem lies in a cleft that is formed by regions of both the protein (DBD and thumb) and the intron RNA (D1d and D4a). Two clusters of amino acids along the third α helix of the protein thumb domain anchor the DNA stem by making contacts at both ends of the DNA helix, at positions separated by approximately one helical turn. The first cluster (S346, R347, R349, R353, N395, and N405) secures the base of the stem through contacts with dG1 and dA2 (Fig. 4C). The second group [S336, Met^{337} (M337), K338, and T339] appears to locally deform the top base pairs of the stem at dC20 and dT21 (Fig. 4C). This is the result of a DNA–protein interaction network that involves insertion of a prolyl-aromatic loop into the distorted, widened minor groove at the tip of the DNA stem. The complementary fit of this peptide loop is mediated by interactions between largely buried side chains [Y278, Phe^{279} (F279) and Pro^{281} (P281)] and the methylene edges of DNA sugar moieties (Fig. 4D). These protein–DNA interactions are supported by contacts between the DNA and RNA backbone residues (dA4O3′ and dG5OP1 with G163 2′OH), which is reminiscent of ribose–zipper interactions that are observed within folded RNA molecules (Fig. 4E) (*15*). Collectively, these interactions enable the holoenzyme to coordinate and selectively identify the shape of a DNA helix.

This shape-selective recognition strategy of the DNA stem is complemented by sequence-specific interactions between the holoenzyme and single-stranded regions of the DNA target. Phylogenetically covarying base pairs are formed between substrate-recognition regions of the intron and single-stranded DNA nucleobases downstream of the DNA stem (*33*). In the holoenzyme, we not only identify these critical WC pairings but also observe a complex network of interactions mediated by the spacer DNA that connects the stem with the IBS sequences. This sequential network of DNA IBS elements and the adjacent spacer

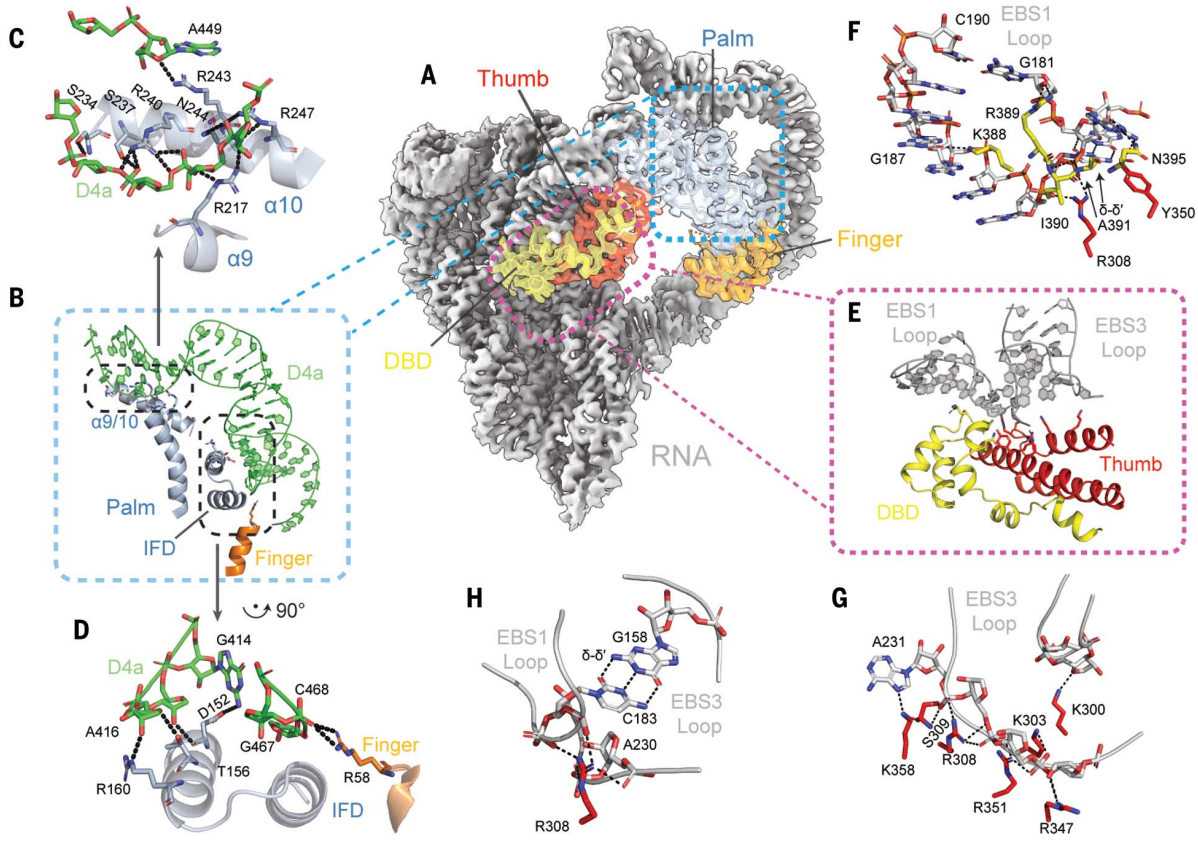


Fig. 3. Mechanism of maturase-facilitated ribozyme catalysis. (**A**) Protein positioning within the retroelement composite map. (**B**) Protein-D4a contact points. (**C** and **D**). Interactions that form the RNA-protein anchor points. (**E**) Protein stabilization of EBS1 and EBS3 loops. (**F** and **G**) Amino acids that rigidify the EBS1 and EBS3 loops. (**H**) R308 joins EBS1 and EBS3 together.

interactions begins with the nucleotide located immediately downstream of the insertion site (dT36), which forms a single–base pair interaction (EBS3-IBS3) with a nucleotide extending from the D1d coordination loop within the intron (A231) (Fig. 4F). Stacked atop this pair is a short helix formed through base pairings between the subsequent stretch of DNA nucleotides (IBS1: dT35, dT34, dT33, and dC32) and a second substrate-recognition loop that projects from the terminus of intron D1d (EBS1: A184, A185, A186, and G187) (Fig. 4F). Similar to the short codon-anticodon helix in the ribosome (*34*), the EBS-IBS1 duplex is further stabilized through the formation of an A-minor motif between A75 and the dC32-G187 base pair (Fig. 4G). The structure reveals that EBS1-IBS1 is not limited to four contiguous base pairs; rather, it is extended by an additional base pair that is formed between the next consecutive nucleotide (A188) in the EBS1 loop and a discontinuous nucleotide from the DNA spacer region (dT30). Indeed, the intervening DNA nucleotide (dT31) is extrahelical and stabilized by interactions with protein residues (Fig. 4H). Through these sequential stacking networks, which are supported by contacts with the protein (i.e., dT36O4 with K361), the intron achieves stable, base-pairing specificity with the DNA target.

Nucleotides within the DNA spacer participate in binding the RNP, adopting an ordered structure that engages in specific interactions with the holoenzyme. Rather than forming a helical stack, the spacer nucleotides (dA28, dT29, dT30, and dT31) form an unusual motif in which the nucleotides splay in alternating directions on either side of the central phosphate spine (Pauling-like DNA), thereby exposing a large interaction interface to the adjacent DBD (Fig. 4, F and H). Amino acids from the DBD intercalate between the DNA spacer nucleotides while forming an abundance of interactions with both the bases and the phosphate backbone (Fig. 4H). For example, N3 of dT31 interacts with amide oxygens of N378, whereas its adjacent phosphate oxygens interact with proximal arginine residues (R381 and R382). Together, these interactions stabilize an unusual backbone conformation that enables the dT30-A188 pair to form atop the EBS-IBS1 helix. In turn, these interactions with the DBD pull the DNA into place, which positions the specialized barb-like structure formed by the α helical bundle within the DBD at the base of the DNA stem (Fig. 4, B and H).

By capping the DNA stem-loop, a set of stacking interactions clamp the loop terminus into position within the holoenzyme. One such interaction forms between the DNA and RNA loop nucleotides that project from D4, which effectively joins the DNA stem and RNA bases into one continuous stacking network. This extended stacking array consists of dA14 and A441 and U443 from D4a (Fig. 4I). This DNA-RNA tertiary interaction is anchored in place by an adjacent stacking network that forms between the extrahelical dT13 residue and a series of conserved amino acid side chains (fig. S8), which form a sequential stack that merges with the hydrophobic core of the protein. The aromatic plane of the dT13 nucleobase and Trp^{136} (W136) flank R266 on either side, which creates an arginine-π stacking sandwich configuration (*35*), in which each component is separated by a planar distance of 3.8 Å (Fig. 4J).

Retroelement primed for attack

To better understand molecular rearrangements that might occur when the intron retroelement binds to DNA substrate, we solved the structure of the apo-RNP, visualizing the free intron-maturase complex at a resolution of

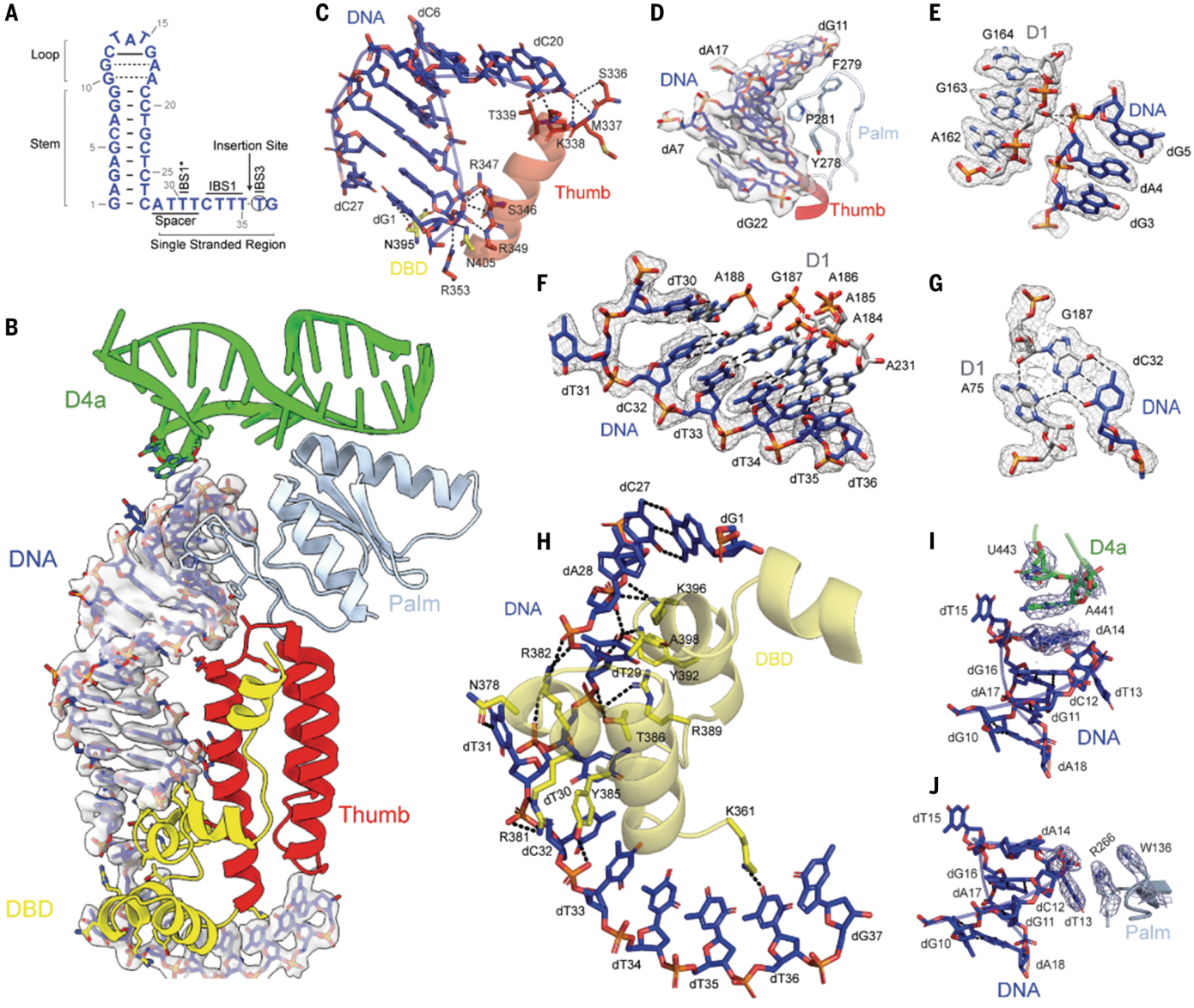


Fig. 4. Shape and sequence recognition of a DNA target. (**A**) Secondary structure of the DNA target. (**B**) Interactions of the structured DNA with holo-RNP. (**C**) Protein contacts with DNA helical stem. (**D**) Fit of the DNA groove against the protein palm linker. (**E**) DNA and D1 backbone interactions. (**F**) EBS-IBS base-pairing interactions. (**G**) Stabilizing A-minor tertiary interaction. (**H**) Interactions between protein and single-stranded DNA. (**I** and **J**) Intermolecular stacking interactions between DNA and (I) RNA nucleotides and (J) protein residues.

3.6 Å (Fig. 5A and fig. S9). We observed that the apo-RNP has an architecture that is almost identical to that of the complex bound to DNA, and that substrate binding induces only minor changes in the structure. The RNP-active site remains completely intact (Fig. 5, A and B) (*9*–*11*). The maturase does not change its orientation in the absence of DNA and remains coordinated at two anchor points along the D4a arm, with the thumb and DBD inserted into the active site to participate in catalysis (Fig. 5A). In this configuration, the binding interface for the target DNA is maintained, which enables the RNP to readily recognize an incoming DNA target and rapidly engage in retrotransposition (Fig. 5A). Upon recognition of the DNA stem-loop, the RNP (palm, fingers, and D4a) appears to become more rigid, as we observe a concomitant increase in local resolution at these positions (figs. S3B and S9B and movies S5 and S6). This is reminiscent of many protein enzymes, whereby docking of ligand into the active site freezes out local motions and locks the substrate in place. In previous ligand-free intron structures, EBS nucleotides were found to be disordered or rearranged (*21*, *22*, *24*). Here, we observe that the positions of the EBS nucleotides are unchanged, likely due to the presence of the maturase. These findings reinforce the mechanistic role of the protein as a stabilizing catalytic component (Fig. 5C). Further evidence of a preformed catalytic core includes the persistence of the heteronuclear metal ion cluster, which remains organized around the lariat, although M2 is not visible in this case (Fig. 5D).

Discussion

Insights into protein-facilitated ribozyme catalysis

The structures presented here reveal how components of the holoenzyme help promote activity of the ribozyme core. The protein buttresses the catalytic residues responsible for specifically

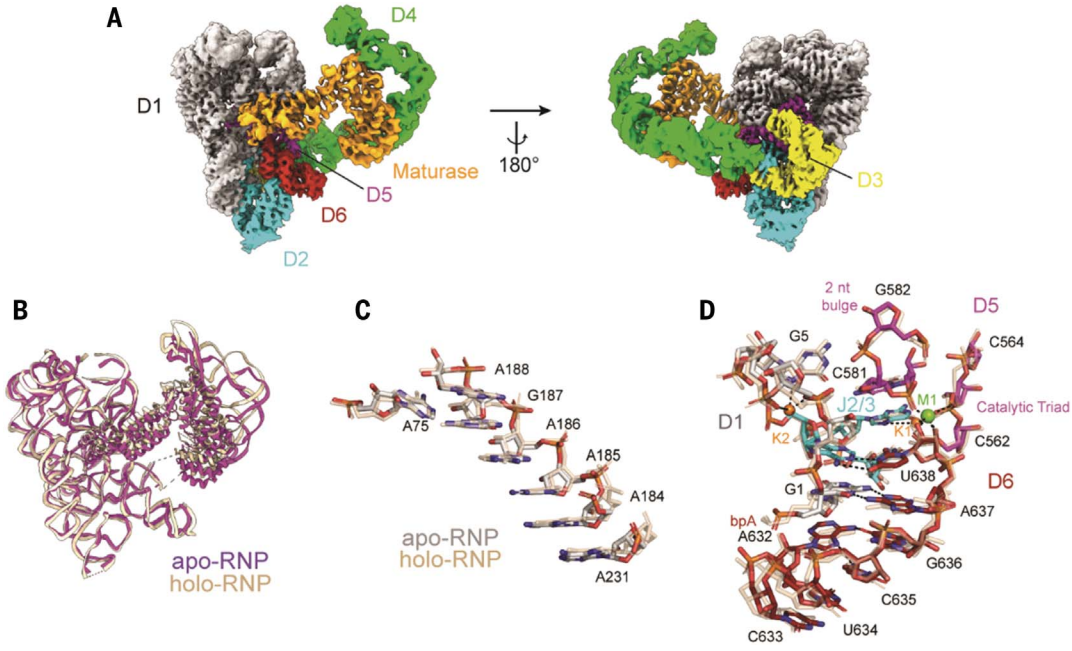


Fig. 5. Retroelement poised to attack. (**A**) Composite cryo-EM map of the apo-RNP. (**B** to **D**) Comparison of (B) the backbone traces, (C) the position of the EBS1 recognition sequences, and (D) the active site of the apo-RNP (colored) and holo-RNP (wheat).

positioning the DNA substrate, providing a missing link in understanding how maturases unlock the full catalytic potential of group II intron ribozymes (Fig. 2). Interactions with protein residues stabilize the intron substrate-recognition loops and precisely arrange nucleotides for DNA binding. These interactions contribute to the proper orientation of reaction components throughout ribozyme catalysis. Our findings provide a direct mechanistic role for the protein cofactor, and they help explain the lowered salt and magnesium requirements in its presence (*36*).

Prior studies, owing to resolution limitations or construct design, were unable to identify a specific function for the maturase protein, except in transient D6 stabilization (*9*). By contrast, we can now show that not only are the protein thumb and DBD proximal to D1 catalytic residues, they have critical roles in stabilizing substrate binding. Given its analogous spatial placement, it is possible that Prp8 may play a similar role during spliceosomal catalysis (fig. S10) (*37*).

RNP recognition of DNA structure: An expanded recognition repertoire

The high-resolution cryo-EM structures we provide here offer a glimpse into RNP strategies for recognizing DNA (Fig. 1B). The holoenzyme structure reveals a stem-loop DNA nestled within the retroelement, bound to RNA and protein. Within this cleft, the protein assists in positioning the insertion site and aligning the DNA stem for steric fit against the complementary maturase surface (Fig. 3). Additional aspects of the unusual recognition strategy include splayed Pauling-like DNA (*38*), a stabilizing A-minor motif, and intermolecular stacking moieties that involve both RNA and protein (Fig. 4). These interactions highlight the symbiotic nature of RNA and protein and underscore the multiplicity of strategies available to RNPs for achieving selective substrate recognition.

The DNA stem-loop motif is exclusive to IIC introns, which contain an abbreviated D1 scaffold and short exon recognition sequences (*7*). In the more highly evolved IIA and IIB introns, the RNP binding motif that we find occupied by the DNA stem in IIC introns is instead replaced by intron insertion motif D1d2, an RNA subdomain that includes EBS2, which is absent in the IIC class (*7*). Comparison of this region across intron classes suggests that EBS2 evolved to imitate the target DNA stem (fig. S11). Indeed, the DNA stem motif structurally resembles the EBS2-IBS2 interactions typical of IIA and IIB introns, and it functionally emulates EBS2 by anchoring the DNA substrate to the RNP. This mimicry suggests that recognition of a structured DNA motif by the more primitive IIC introns was replaced by RNA domains within the intron itself, which resulted in longer target-recognition sequences that provided greater base-pairing specificity for the retroelement.

Implications for reverse splicing and reverse transcription

Encircled by RNA, the exterior surfaces of the protein are enclosed, but the concave interior of the protein, adjacent to the catalytic core, is conspicuously solvent accessible, which has functional implications. During reverse splicing, the D6 helix undergoes a conformational change that places the lariat linkage into the active site (*9*). To accomplish this, D6 disengages from D2 and swings 90° upward, contacting D1c and a basic patch on the protein thumb. Our structures do not preclude D6 helix dynamics, because there is ample space for a similar movement and the regions that D6 contacts are accessible. The open architecture we observe provides a direct route for DNA to approach the RT active site, because it remains unobstructed by other intron domains and can readily accommodate an entire hybrid duplex for reverse transcription (*39*). This suggests that initiation of RT activity, within the current holoenzyme assembly, may be possible without marked conformational rearrangement.

Retroelement poised to attack

Group II intron retroelements are proliferative, invasive agents, and our structures explain why. The apo-retroelement is poised to react and does not require any reorganization of structure upon target DNA binding. The arrangement of the active site, from substrate-recognition nucleotides to the heteronuclear metal ion cluster to the DNA binding interface, is preserved despite the absence of DNA substrate (Fig. 5). This prearranged organization is consistent with the biological role of group II introns as parasitic genetic elements (*40*). Use of the same catalytic core from splicing to integration eschews the need for major rearrangements or host cofactors and allows complete autonomy, which is highly advantageous for a genetic parasite.

Total integration of the RNP requires faithful and accurate reverse transcription of the intron sequence, including the long open reading frame (ORF) that encodes the protein, after insertion. This is accomplished by using

the RT activity of the multifunctional maturase. MarathonRT, the protein within the holoenzyme visualized here, is a well-characterized, robust, accurate, and ultraprocessive RT enzyme that is capable of copying through long, structurally complex templates (*41*). The intimate association of the parent intron with this protein allows access to its exceptional RT properties and ensures that the intron sequence, which is pivotal to its tertiary architecture, is preserved, allowing the retroelement to continually propagate.

Implications for modern retroelements

Study of group II intron complexes provides a window into our understanding of non-LTR retrotransposons, such as the L1 RNP, an active mobile element that continues to disperse in human genomes (*42*, *43*). Computationally predicted structures of ORF2p, the mobility factor of L1, show that its RT and thumb domain resemble that of the maturase, MarathonRT (fig. S12) (*44*). ORF2p contains an additional N-terminal endonuclease and a C-terminal extension, but these domains do not block the exterior basic surfaces of the RT and thumb. MarathonRT and ORF2p are evolutionarily related, and they are implicated in similar mobility mechanisms, so it is possible that the same surfaces are used for anchoring and substrate recognition (*45*). Given the lack of structural information on L1 and the strong parallels between systems, our work provides a starting point for imagining how L1 might assemble and function.

This study reveals strategies for RNP interactions with DNA, having implications for mechanistic understanding of the spliceosome and non-LTR retrotransposons.

REFERENCES AND NOTES

1. T. H. Eickbush, *Curr. Biol.* **9**, R11–R14 (1999).
2. J. L. Ferat, F. Michel, *Nature* **364**, 358–361 (1993).
3. L. Bonen, J. Vogel, *Trends Genet.* **17**, 322–331 (2001).
4. A. M. Lambowitz, S. Zimmerly, *Cold Spring Harb. Perspect. Biol.* **3**, a003616 (2011).
5. H. Wank, J. SanFilippo, R. N. Singh, M. Matsuura, A. M. Lambowitz, *Mol. Cell* **4**, 239–250 (1999).
6. S. Zimmerly, H. Guo, P. S. Perlman, A. M. Lambowitz, *Cell* **82**, 545–554 (1995).
7. A. M. Pyle, *Crit. Rev. Biochem. Mol. Biol.* **45**, 215–232 (2010).
8. A. M. Pyle, *Annu. Rev. Biophys.* **45**, 183–205 (2016).
9. D. B. Haack *et al.*, *Cell* **178**, 612–623.e12 (2019).
10. G. Qu *et al.*, *Nat. Struct. Mol. Biol.* **23**, 549–557 (2016).
11. N. Liu *et al.*, *Nucleic Acids Res.* **48**, 11185–11198 (2020).
12. L. Dai, N. Toor, R. Olson, A. Keeping, S. Zimmerly, *Nucleic Acids Res.* **31**, 424–426 (2003).
13. C. Zhao, A. M. Pyle, *Nat. Struct. Mol. Biol.* **23**, 558–565 (2016).
14. C. Zhao, F. Liu, A. M. Pyle, *RNA* **24**, 183–195 (2018).
15. N. Toor, K. S. Keating, S. D. Taylor, A. M. Pyle, *Science* **320**, 77–82 (2008).
16. R. T. Chan, A. R. Robart, K. R. Rajashankar, A. M. Pyle, N. Toor, *Nat. Struct. Mol. Biol.* **19**, 555–557 (2012).
17. A. M. Lentzsch, J. L. Stamos, J. Yao, R. Russell, A. M. Lambowitz, *J. Biol. Chem.* **297**, 100971 (2021).
18. L. Dai, S. Zimmerly, *Nucleic Acids Res.* **30**, 1091–1102 (2002).
19. N. Toor, A. R. Robart, J. Christianson, S. Zimmerly, *Nucleic Acids Res.* **34**, 6461–6471 (2006).
20. C. Zhao, A. M. Pyle, *Trends Biochem. Sci.* **42**, 470–482 (2017).
21. A. R. Robart, R. T. Chan, J. K. Peters, K. R. Rajashankar, N. Toor, *Nature* **514**, 193–197 (2014).
22. M. Costa, H. Walbott, D. Monachello, E. Westhof, F. Michel, *Science* **354**, aaf9258 (2016).
23. S. M. Fica, M. A. Mefford, J. A. Piccirilli, J. P. Staley, *Nat. Struct. Mol. Biol.* **21**, 464–471 (2014).
24. M. Marcia, A. M. Pyle, *Cell* **151**, 497–507 (2012).
25. S. M. Fica *et al.*, *Nature* **503**, 229–234 (2013).
26. M. E. Wilkinson, S. M. Fica, W. P. Galej, K. Nagai, *Mol. Cell* **81**, 1439–1452.e9 (2021).
27. O. Fedorova, A. M. Pyle, *EMBO J.* **24**, 3906–3916 (2005).
28. B. Cousineau, S. Lawrence, D. Smith, M. Belfort, *Nature* **404**, 1018–1021 (2000).
29. N. Toor, G. Hausner, S. Zimmerly, *RNA* **7**, 1142–1152 (2001).
30. R. N. Singh, R. J. Saldanha, L. M. D'Souza, A. M. Lambowitz, *J. Mol. Biol.* **318**, 287–303 (2002).
31. C. Zhao, A. M. Pyle, *Curr. Opin. Struct. Biol.* **47**, 30–39 (2017).
32. M. Matsuura, J. W. Noah, A. M. Lambowitz, *EMBO J.* **20**, 7259–7270 (2001).
33. F. Michel, K. Umesono, H. Ozeki, *Gene* **82**, 5–30 (1989).
34. A. Lescoute, E. Westhof, *Biochimie* **88**, 993–999 (2006).
35. M. M. Flocco, S. L. Mowbray, *J. Mol. Biol.* **235**, 709–717 (1994).
36. M. Matsuura *et al.*, *Genes Dev.* **11**, 2910–2924 (1997).
37. S. M. Fica, C. Oubridge, M. E. Wilkinson, A. J. Newman, K. Nagai, *Science* **363**, 710–714 (2019).
38. L. Pauling, R. B. Corey, *Proc. Natl. Acad. Sci. USA* **39**, 84–97 (1953).
39. J. L. Stamos, A. M. Lentzsch, A. M. Lambowitz, *Mol. Cell* **68**, 926–939.e4 (2017).
40. A. M. Lambowitz, S. Zimmerly, *Annu. Rev. Genet.* **38**, 1–35 (2004).
41. L. T. Guo *et al.*, *J. Mol. Biol.* **432**, 3338–3352 (2020).
42. C. R. Beck, J. L. Garcia-Perez, R. M. Badge, J. V. Moran, *Annu. Rev. Genomics Hum. Genet.* **12**, 187–215 (2011).
43. M. A. Kerachian, M. Kerachian, *Clin. Chim. Acta* **488**, 209–214 (2019).
44. J. Jumper *et al.*, *Nature* **596**, 583–589 (2021).
45. Y. Xiong, T. H. Eickbush, *EMBO J.* **9**, 3353–3362 (1990).

ACKNOWLEDGMENTS

We thank M. Llaguno, S. Wu, J. Lin, K. Zhou, and K. Gibson (YCRC) for help with grid preparation, sample screening, and data collection. We thank F. Bleichert for helping with cryoSPARC data processing and Y. Xiong, K. Zhang, and C. Wang for helpful suggestions. We also thank C. Zhao and O. Federova for insights and advice throughout this project. **Funding:** This work was supported by the Howard Hughes Medical Institute and the Gruber Foundation (Gruber Science Fellowship to K.C.). Cryo-EM data were collected with microscopes at the Yale CryoEM Resource Core that is funded in part by the NIH (S10OD023603). Funding for open access charge was provided by the Howard Hughes Medical Institute. A.M.P. is an investigator and L.X. is a research associate with the Howard Hughes Medical Institute. **Author contributions:** K.C. and L.X. designed the protocol to purify the retroelement complexes, prepared the samples, made EM grids, and performed biochemical assays. S.C.D. conducted SEC-MALS experiments on purified intron complexes. K.C. and L.X. collected EM data. K.C. and L.X., with assistance from P.C. and S.C.D., processed the EM data. K.C. and L.X., with help from J.P. and S.C.D., built the atomic model. K.C., L.X., and A.M.P. analyzed the structure. K.C. and L.X. drafted and prepared the manuscript. A.M.P. supervised and coordinated the group II intron project. **Competing interests:** A patent application on MarathonRT has been filed by Yale University. **Data and materials availability:** All data are available in the main text and the supplementary materials. Cryo-EM maps are available in the Electron Microscopy Data Bank with codes EMD-26550 (holo-RNP) and EMD-26549 (apo-RNP). Structural models are available in the Protein Data Bank with PDB accession codes 7UIN (holo-RNP) and 7UIM (apo-RNP). **License information:** Copyright © 2022 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abq2844
Materials and Methods
Figs. S1 to S12
Table S1
References (*46*–*65*)
MDAR Reproducibility Checklist
Movies S1 to S6

View/request a protocol for this paper from *Bio-protocol*.

Submitted 29 March 2022; resubmitted 10 June 2022
Accepted 17 October 2022
10.1126/science.abq2844

REVIEW SUMMARY

RADIO ASTRONOMY

The discovery and scientific potential of fast radio bursts

Matthew Bailes

BACKGROUND: Fast radio bursts (FRBs) are broadband, millisecond-duration bursts of radio emission visible at cosmological distances. As FRBs traverse the intergalactic medium, their radiation is slightly delayed by the presence of free electrons in a frequency-dependent manner. When these delays are combined with cosmological models of the distribution of baryons in the Universe, they can infer distances. The first FRB, the Lorimer Burst, was discovered in 2007. Although its radio brightness was similar to that of radio pulsars (neutron stars) in the Milky Way, its inferred distance was a million times greater, indicative of a new class of object.

Instrumental and computational advances made FRB discovery routine by the mid-2010s, and there are many thousands of bursts known from more than 600 unique sources. Some FRBs arise from sources that produce multiple bursts, separated by anything from seconds to months, and these are known as "repeaters." However, the majority of FRB sources have never been seen to repeat.

Because most of the baryonic (normal) matter in the Universe is ionized, FRBs can constrain the total baryonic content of the Universe. The discovery of FRBs raised two scientific questions: (i) what produces them? and (ii) what can they tell us about the Universe?

ADVANCES: The early FRBs were discovered using single radio dishes with limited spatial resolution. These instruments established that FRBs might be detectable to redshifts corresponding to when the Universe was only half its current age. However, they were unable to localize any FRBs to their host galaxies, so their true distances remained uncertain. In 2016, the first repeating FRB was discovered, which allowed follow-up observations using radio interferometers with better spatial resolution. These showed that the repeating source is situated in a small, low-metallicity dwarf galaxy at a distance confirming their cosmological nature and near a persistent source of radio emission. The localization of the repeater thus demonstrated that the luminosities of FRBs were extremely high.

Improved instrumentation greatly expanded both the number of observed FRBs and the number with identified host galaxies. Localizations of repeaters determined numerous host galaxies and showed that repeating FRBs are likely to be associated with young, highly magnetic neutron stars (magnetars). Some repeaters appear to have cyclic activity windows, which is consistent with an orbit, or possibly a precession, of the source. The degree of linear polarization of repeaters is strongly frequency dependent, indicating that they are often located in highly magnetic, ionized environments.

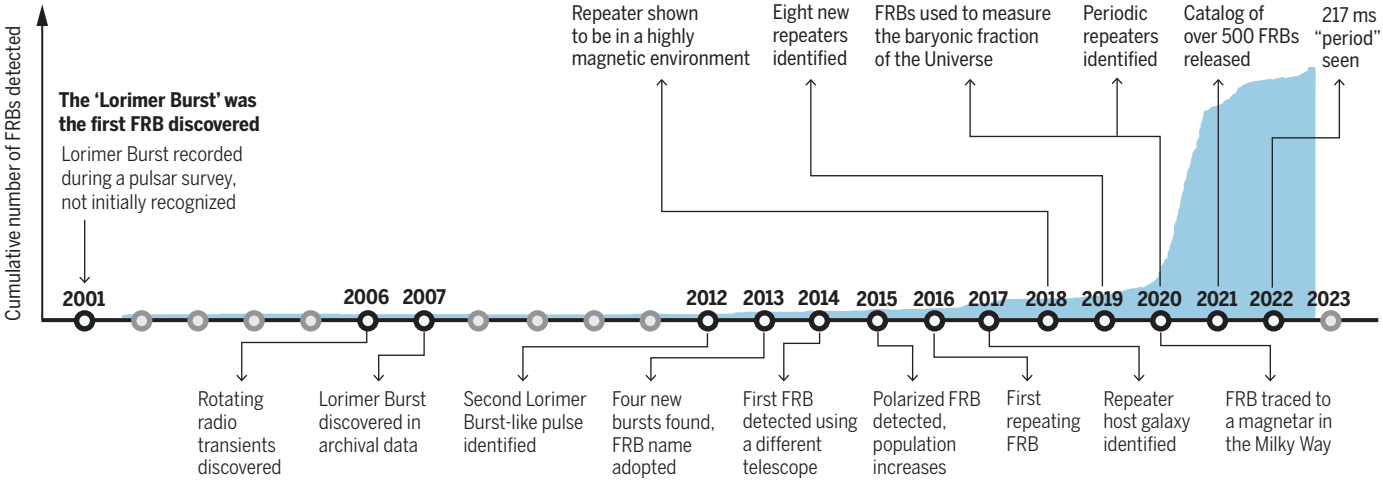
Further developments in instrumentation enabled the localization of many nonrepeating FRBs. Combining all the known FRB host galaxies led to a measurement of the baryonic content of the Universe. More than 1000 FRBs were detected from a single repeater but with no underlying periodicity. In 2020, an FRB was observed from a magnetar within the Milky Way.

OUTLOOK: The diversity of FRBs continues to expand. One FRB had emission extended over several seconds, punctuated by bursts with a 217-ms periodicity. Others showed weaker evidence for faster periodicities, at 2.8 and 10.7 ms, which could be linked to the rotation period of the neutron stars thought to produce them. The initial 12 orders of magnitude in luminosity between the early FRBs and pulsars in the Milky Way is being closed by observations of fainter FRBs. The nature of the sources and the emission mechanism remains unclear. It is possible that there are multiple ways of producing FRBs, including from unusual locations such as millisecond pulsars in globular clusters.

New instruments are currently being constructed and commissioned that will have increased sensitivity. These will also localize many more FRBs to their host galaxies, increasing the utility of FRBs in cosmology. ■

The list of author affiliations is available in the full article online.
Email: mbailes@swin.edu.au
Cite this article as M. Bailes, *Science* **378**, eabj3043 (2022). DOI: 10.1126/science.abj3043

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abj3043



Timeline of some important breakthroughs in FRBs. The blue graph indicates the cumulative number of FRBs detected (~800, including some bursts from repeaters). Source: HeRTA: FRBSTATS online catalog.

REVIEW

RADIO ASTRONOMY

The discovery and scientific potential of fast radio bursts

Matthew Bailes

Fast radio bursts (FRBs) are millisecond-time-scale bursts of coherent radio emission that are luminous enough to be detectable at cosmological distances. In this Review, I describe the discovery of FRBs, subsequent advances in understanding them, and future prospects. Thousands of potentially observable FRBs reach Earth every day, which likely originate from highly magnetic and/or rapidly rotating neutron stars in the distant Universe. Some FRBs repeat, with this subclass often occurring in highly magnetic environments. Two repeating FRBs exhibit cyclic activity windows, consistent with an orbital period. One nearby FRB was emitted by a Galactic magnetar during an x-ray outburst. The host galaxies of some FRBs have been located, providing information about the host environments and the total baryonic content of the Universe.

Fast radio bursts (FRBs) are millisecond flashes of radio waves from distant astronomical sources. The first FRB (*1*), later named the Lorimer Burst, was discovered in 2007, some 40 years after the discovery of radio pulsars (*2*). As was the case with pulsars, its discovery was completely serendipitous. The burst swept across the 288-MHz passband of the radio receiver in about a third of a second, consistent with emission from a distant extraterrestrial source, with the radio waves subsequently dispersed by passage through an ionized plasma (Fig. 1, D and E). It struck the Parkes 64-m radio telescope in Australia just after 5:50 a.m. local time on 25 August 2001, and was automatically archived onto magnetic tape as part of a radio pulsar survey. It remained undiscovered for >5 years. This Review examines its serendipitous discovery, the subsequent demonstration that it was part of a previously unknown class of astronomical source, and the cosmological implications.

Discovery of the Lorimer Burst

The discovery of FRBs became possible because of the development of high-time-resolution radio instrumentation and software tools, which are mainly used for pulsar surveys. Radio pulsars are rapidly rotating, highly magnetized neutron stars that emit beams of emission, which appear to pulse as the neutron star rotates (*2*–*5*) in a manner similar to a lighthouse. By 2007, more than half of the known pulsars had been discovered using the Parkes 64-m telescope (Fig. 2A) as a result of its low radio interference environment and access to the Southern Hemisphere sky, where most pulsars reside (*6*–*9*).

Centre for Astrophysics and Supercomputing, Swinburne University of Technology, Hawthorn, Victoria 3122, Australia. Email: mbailes@swin.edu.au

The study of radio pulsars has been advanced by discoveries of unusual objects, usually in large-scale surveys (*2*, *10*, *11*). A new class of pulsar-like objects, the rotating radio transients (RRATs) were discovered in 2006 using the Parkes Multibeam Receiver (*12*), a 13-pixel radio camera that was used in multiple pulsar surveys (*6*, *7*, *13*, *14*). RRATs (*15*) were initially interpreted as an atypical type of pulsar that only rarely emitted pulses, albeit always in phase with a neutron star's rotation period. This was unlike radio pulsars, which usually emitted regular pulsations with each rotation. In 2007, searches were underway to find more examples of RRATs (*16*–*18*).

Like pulsars, RRATs exhibit a radio frequency–dependent delay that appears as a sweep on the radio receiver. This arises because radio waves travel slightly slower in the ionized interstellar medium than the speed of light in a vacuum (c). The speed is determined by the radio frequency (ν) and the density of free electrons (Fig. 1A). By the time a broadband radio pulse arrives at Earth, this frequency-dependent delay leads to a well-defined sweep in which the higher radio frequencies arrive before the lower ones (shown in Fig. 1C for PSR J1707–4053). A radio pulse's dispersion measure, $DM \equiv \int_0^L n_e dL$, is the integrated column density of free electrons along the line of sight, where n_e is the local free electron density in cubic centimeters, L is the distance to the source in parsecs (pc), and DM is the dispersion measure expressed as parsecs per cubic centimeter. The difference in the arrival times ($t_2 - t_1$) between two radio waves of frequencies ν_2 and ν_1 is approximately as follows (*19*):

$$t_2 - t_1 \approx 4.15\left[\left(\frac{1}{\nu_2^2}\right) - \left(\frac{1}{\nu_1^2}\right)\right] DM \text{ ms} \quad (1)$$

where the radio frequency is expressed in gigahertz. The measured DM of a source, combined with a free electron model, can be used as a proxy for distance within the Galaxy (*20*, *21*) and for cosmological distances (*22*). Although it greatly complicates radio pulsar instrumentation and analysis, pulse dispersion helps observers distinguish celestial sources from local radio interference, and is ultimately the effect that allows FRBs to have cosmological applications.

D. R. Lorimer and undergraduate student D. J. Narkevic searched for RRATs in an archival multibeam pulsar survey of the Small Magellanic Cloud (SMC), a nearby dwarf galaxy, using the Parkes 64-m telescope. Narkevic's initial analysis had detected a total of two putative RRATs in the whole SMC survey. They both appeared at exactly the same time with DM ~375 pc cm^{-3} in two adjacent beams of the 13-beam receiver, suggesting that either a very bright source was illuminating the sidelobes of multiple beams or there was an unusual form of radio interference (Fig. 1B).

I was observing at the Parkes telescope on an unrelated project with Lorimer and became involved in diagnosing the putative dispersed radio source. We extracted the relevant burst data and displayed them as waterfall plots, which show the pulse energy as a function of both time and frequency (*23*, *24*). We found that the source had a dispersion sweep similar to those of pulsars, with evidence of radio frequency–dependent multipath interstellar scattering (which scales as ν^{-4}), as is often seen in observations of pulsars (*25*) (Fig. 1, D and E). Further investigation showed that the burst had saturated the receiver in the beam closest to the source's location (beam 6 in Fig. 1B), which have limited dynamic range, causing an algorithm designed to remove interference to replace the burst with synthetic data. If the instrument had not been sensitive to strong signals in multiple beams, then the burst would never have been found. Disabling the interference rejection algorithm and reprocessing the archived data showed that the burst was >100 times the survey's detection threshold, with an estimated flux density of 30 jansky (Jy) (Fig. 1, D and E). The burst was bright enough to be visible in four of the 13 beams (*26*, *27*).

The estimated source distance placed it well beyond the SMC (the target of the survey); it appeared to be at cosmological distance (*1*). In the voids between galaxies, free electron densities (*22*) are ~1 m^{-3}. This makes the source distance equivalent to ~1 Mpc for each 1 pc cm^{-3} of the DM after removing the Milky Way foreground. Thus, the DM of ~375 pc cm^{-3} indicated a distance of ~1 Gpc; for comparison, the Milky Way is ~30 kpc across and the SMC is at a distance of ~60 kpc. To explain the observed brightness at such a distance, the source

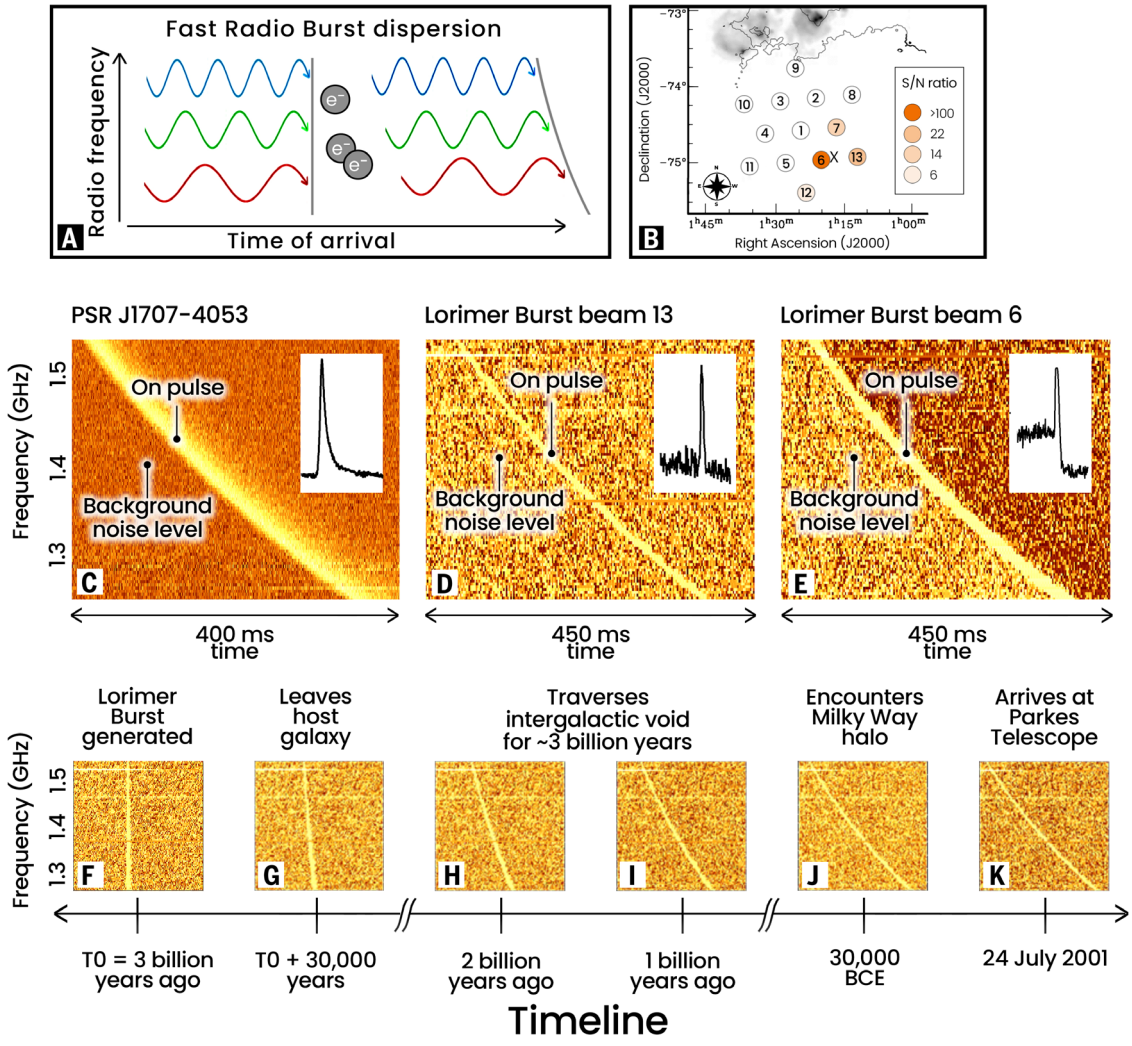


Fig. 1. FRB dispersion and the location of the Lorimer Burst. (**A**) Conceptual illustration of how dispersion delays the time of arrival at Earth. As radio waves encounter free electrons, they become delayed in a radio frequency–dependent manner. The more energetic (higher-frequency) radio waves (blue) experience less delay than the lower energy waves (red). This leads to a characteristic sweep observed in FRBs and pulsars. (**B**) Pointing of the Parkes 13-beam receiver just south of the edge of the SMC at the time of the observation of the Lorimer Burst (*1*). The burst-saturated beam 6, was well above the detection threshold in beams 7 and 13, and was weakly detected in beam 12 (*27*). The cross indicates the inferred burst position (*1*). (**C**) Observed dispersion sweep of the pulsar PSR J1707–4053 with DM = 360 pc cm^{-3} and its de-dispersed pulse profile (inset) (*125*). (**D** and **E**) Dispersion sweep and integrated pulse profile (inset) of the Lorimer burst at DM = 375 pc cm^{-3} in sidelobe beam 13 (D) and beam 6 (E). The dip in flux after the burst in beam 6 is an instrumental artifact caused by saturation. (**F** to **K**) Inferred evolution of the Lorimer Burst's dispersion over cosmic time. T_0 is the time of emission.

would have to be about a trillion times more luminous than any known pulsar. Alternatively, the source could have been enshrouded in a highly ionized plasma in its host environment, leading to a spurious distance estimate. There is a subclass of high-magnetic-field and/or short-period pulsars that are known to occasionally emit highly energetic radio pulses, which might have been linked to the burst. The Crab Pulsar, a young and highly magnetized neutron star, occasionally produces these giant pulses, individual radio flashes that can be much brighter than the mean energy of its average pulse. However, even the brightest Crab Pulsar giant pulses were about a trillion times less energetic than the putative burst (Fig. 3). The implied radio energy emitted in the 5-ms-duration burst was similar to all the power the Sun emits over a month.

The burst was well above the survey detection threshold. In extragalactic surveys, a homogeneous isotropic cosmological population follows the relation d log N/d log S = –3/2, where N is the number of sources above a flux density S. This arises because the volume surveyed expands as the distance D^3, whereas each object's flux density follows the inverse square law D^{-2}. For every Lorimer Burst, there should be many (perhaps dozens) of fainter bursts present in the survey data. There did not appear to be any. [A FRB search (*28*) found another FRB in the same dataset, with a DM of 1187 pc cm^{-3}, but that was only identified years later.]

If the source repeated, then it would provide more confidence in its celestial origin, but 40 hours of follow-up observations at Parkes saw nothing (*1*). It appeared that the burst was either (i) a one-off hyperluminous flash of radio waves from the distant Universe a trillion times more luminous than known radio transient bursts and requiring improbable survey statistics or (ii) some obscure form of radio interference. We decided that the pulse's well-defined sweep with frequency was sufficient to conclude that it was celestial, so we submitted a paper summarizing our findings (*1*). The event soon became known

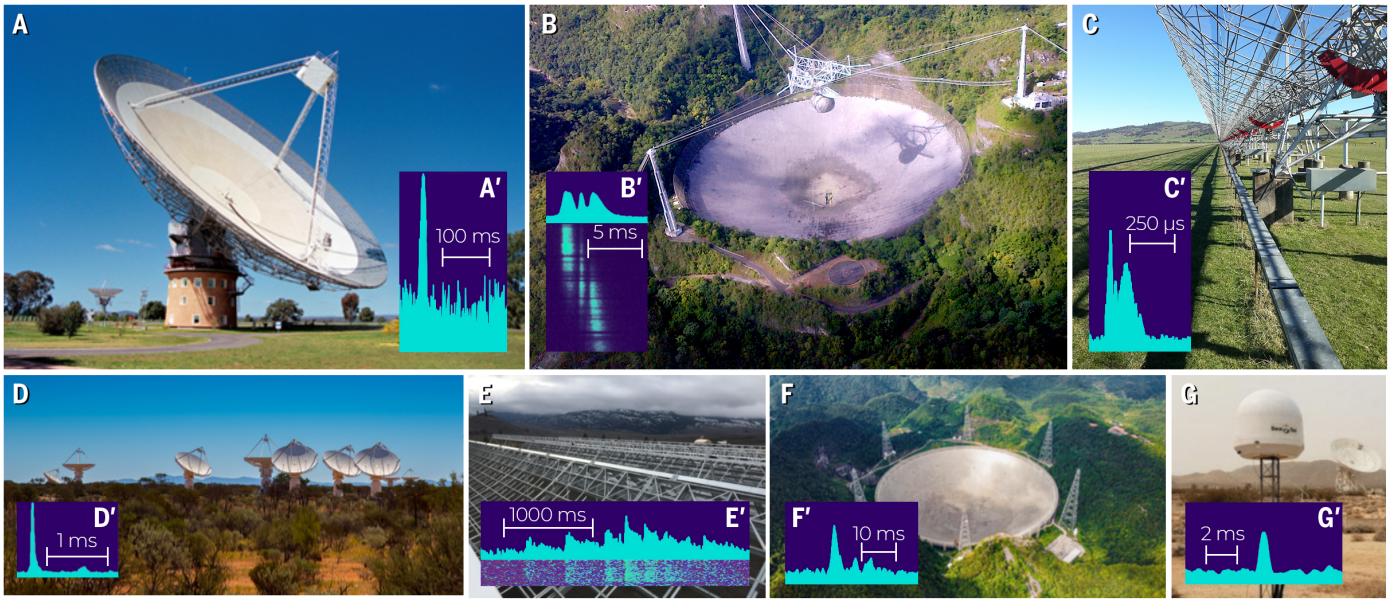


Fig. 2. Seven radio telescopes and an example FRB detected at each telescope. Insets show the intensity of each burst as a function of time (integrated profiles). FRBs with complex frequency structure [(B′) and (E′)] also have waterfall plots beneath their integrated profiles. (**A**) The Parkes 64-m telescope and the Lorimer Burst FRB 010724 (*1*) (inset). (**B**) The 305-m Arecibo Observatory and one burst from the repeating FRB 121102 (*73*). The burst exhibits a downward-drifting frequency effect as a function of time. Frequencies run from low to high in the vertical direction. (**C**) The 1.5-km-long Molonglo telescope and FRB 170827, which has an extremely narrow temporal structure observed using the UTMOST real-time data capture system (*77*, *126*). (**D**) The core of the 36-antenna ASKAP telescope. The inset shows the four-component FRB 181112, which has narrow temporal features (*78*). (**E**) The cylindrical 4 × 20 m wide × 100 m long CHIME telescope and a 3-s-long FRB 20191221A that exhibited a 216.8-ms periodicity (*88*). (**F**) The 500-m FAST telescope and the three-component FRB 181123 (*127*). (**G**) The STARE2 telescope and the Galactic FRB 200428, which was emitted by a magnetar during an x-ray flare (*86*). [Photo credits: J. Sarkissian (A), F. Camilo (B), C. Flynn (C), K. Steele (D), M. Bailes (E), Di Li (F), and S. R. Kulkarni (G).]

as the "Lorimer Burst," now also designated FRB 010724.

Implications of the Lorimer Burst

The apparent luminosity of the Lorimer Burst was very high, but there were few clues as to what may have caused it. There was only an upper limit (≤5 ms) on the intrinsic width of the burst due to a combination of instrumental broadening and radio wave scattering. Requiring the emission region to be causally connected (within 5 ms of travel time at the speed of light) set an upper limit on its size of ≤1500 km. Such distance scales are consistent with the dimensions of the rotating magnetic fields seen in rapidly rotating neutron stars, shocks from explosions emanating from relativistic objects, or collisions between neutron stars and other compact objects (neutron stars or stellar-mass black holes).

Radio telescopes observe a small fraction of the sky, so initial estimates of the FRB event rate were hundreds per day for FRBs as bright as the Lorimer Burst (*1*) and 10,000 d^{-1} for fainter bursts that were still above the detection limit of large radio dishes. The implied rate in a given cosmological volume was similar to that of supernovae: about once every few decades in a normal galaxy.

If its origin could be determined, then the Lorimer Burst had potential for use as a cosmological probe. The *DM* contains information on the number of free electrons along its path. Most of the Universe's baryonic (normal atomic) mass is not in galaxies, but between them in the intergalactic medium (IGM). The bulk of the IGM mass is in hydrogen and helium, which do not retain their electrons because they are ionized by ultraviolet light. An FRB can therefore act as a free electron (and hence baryon) counter between the host galaxy and Earth (Fig. 1, F to K). Localized FRBs could potentially constrain the total mass of IGM, a quantity with controversial measurements using other methods (*29*).

The hunt for more FRBs

Attempts were made to find more bursts, both from existing archival data and by initiating surveys that specifically targeted FRBs. However, none of the early efforts was successful. Many groups searched archival data, finding some additional RRATs (*18*, *30*), but no events similar to the Lorimer Burst were found. Surveys specifically designed to find new bursts (*31*, *32*) also initially did not detect any. It was discovered that the Parkes Radio Telescope receiver was often struck by quasidispersed radio pulses that were present in all 13 beams of the multibeam receiver, some of which had dispersion similar to the Lorimer Burst (*33*). These were shown to be caused by a microwave oven ~100 m from the Parkes dish (*34*), raising further doubts. Was the Lorimer Burst just a similar form of radio interference?

In 2012, another search of the Parkes Multibeam Pulsar Survey found a potential Lorimer Burst–like signal (*35*) with a *DM* of 745 pc cm^{-3}, although it was uncertain whether it originated outside of the Galaxy (*36*). The celestial and cosmological nature of these dispersed radio signals was therefore uncertain.

2013: A cosmological population

In 2008, three pulsar and radio burst surveys commenced at Parkes, called the High Time Resolution Universe (HTRU) surveys (*37*). These used multibeam anti-coincidence detection methods (*38*) that proved effective at removing terrestrial near-field interference. In 2013, one of the HTRU surveys found four radio bursts (*39*), the brightest of which had a *DM* almost three times that of the Lorimer Burst that followed the same ν^{-2} dispersion and ν^{-4} scattering power laws with radio frequency. This demonstrated that fainter and higher-dispersion (more distant) bursts existed, implying that the Lorimer Burst was part of a cosmological population. The term "fast radio burst" and the acronym FRB were coined at that time. A system of nomenclature was adopted, designating each burst FRB followed by numerals indicating the date it

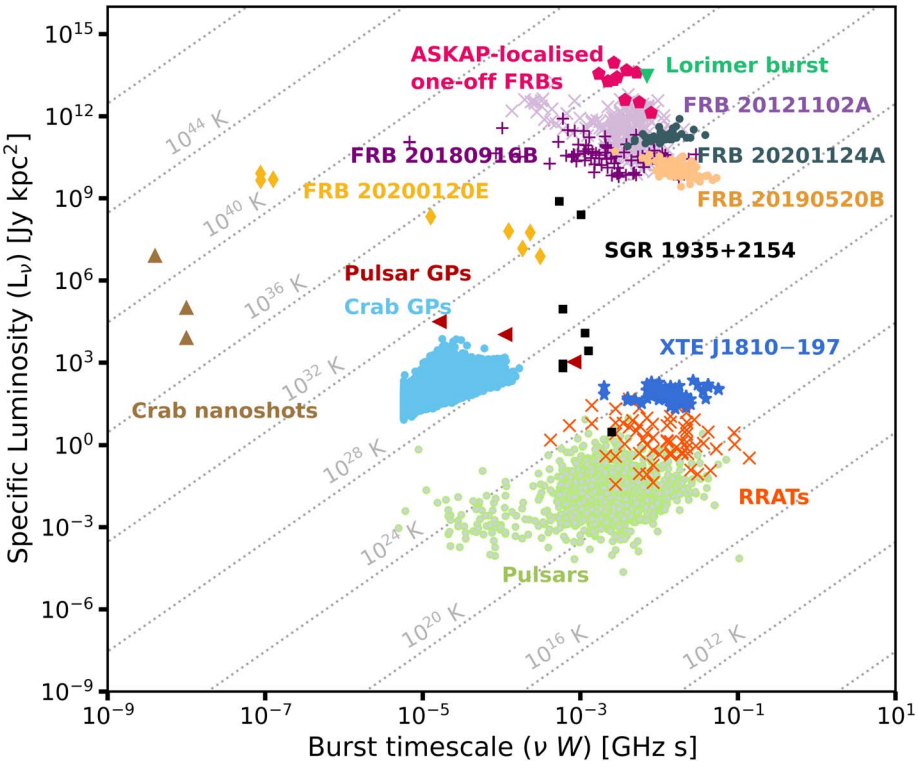


Fig. 3. Luminosity as a function of burst time scale for short-duration coherent radio emitters. The Lorimer Burst (green triangle) is a trillion times more energetic than radio pulsars (light green circles) and RRATs (orange crosses) within the Milky Way. Also shown are other one-off FRBs (magenta pentagons), repeating bursts from FRB 20121102A (light purple crosses) and FRB 20180916B (dark purple plus symbols), the repeating FRB 20200120E in a globular cluster of M81 (yellow diamonds), and radio bursts from the galactic magnetars SGR 1935+2154 (black squares) and XTE J1810−197 (blue stars). Nanosecond duration bursts from the Crab Pulsar (brown triangles) have similar brightness temperatures (the black body temperature that would produce equivalent brightness) to the most energetic FRBs. The sloped dotted gray lines indicate the brightness temperature, which is proportional to the specific intensity of a source. Luminosities and time scales are from (*1*, *76*, *128*).

occurred, similar to that used for for gamma-ray bursts. For example, FRB 110220 was detected on 20 February 2011 Universal Coordinated Time (UTC); it had a dispersion measure of 995 pc cm^{-3} (*39*). Models of extragalactic dispersion indicated that the sources could be at distances up to redshift $z \sim 1$, when the Universe was half its current age.

2014–2017: Single-dish discoveries

Although the four additional FRBs were reassuring, they had also been discovered with the Parkes 64-m telescope. Was there something in its local environment that was mimicking dispersed pulses? This doubt was dispelled in 2014, when a team using the Arecibo 305-m dish in Puerto Rico (Fig. 2B) announced the detection of another FRB (*40*). FRB 121102 had a *DM* of 556 pc cm^{-3} and, like the Parkes examples, appeared to be far beyond the Milky Way. Another *DM* = 790 pc cm^{-3} FRB was found in archival data (*41*), and then a bright FRB was identified using the 100-m Green Bank Telescope (GBT) (*42*), which recorded full polarimetric information showing that the FRB had a polarization fraction of almost 50%. Changes in the position angle of the polarization as a function of frequency (known as Faraday rotation) were evident, indicating that the FRB source was probably immersed in a highly magnetized region within its host galaxy. Another five FRBs (*43*) were identified in the HTRU survey, including evidence that the bursts might have multiple components, and a *DM* as high as 1629 pc cm^{-3} was found. Other FRBs well above the detection threshold were also discovered using the Parkes 64-m telescope (*44*, *45*).

Single-dish radio telescopes have poor spatial resolution (several square degrees), so none of these FRBs could be unambiguously associated with a host galaxy. A definitive demonstration that FRBs are at cosmological distances could potentially be made by using an interferometer, which has much better spatial resolution, to localize an FRB to a host galaxy (*46*). Without known distances or association at other wavelengths, it was difficult to constrain models of emission mechanisms or even be certain that FRBs were a cosmological population.

Early physical models

A catalog of theories for the emission mechanism of FRBs (*47*) listed more models than there were then known bursts. If FRBs are at cosmological distances, then their estimated radio energies are 10^{38} to 10^{40} ergs. This is about the same total energy as the Sun emits in a day to a month, but all in the radio band and all within a few milliseconds. Although this eliminates Sun-like stars as the source of FRBs, there are nevertheless many potential astrophysical sources. Accreting neutron stars often exhibit x-ray luminosities of 10^{38} ergs s^{-1} and the Crab Pulsar releases its rotational kinetic energy at a rate of 4×10^{38} ergs s^{-1}. Either could regularly emit a low-powered FRB without violating energy conservation.

Causality requires that the dimension (d) of an FRB source must be $d \leq c\delta t$, where c is the speed of light and δt is the duration of the FRB. For observed FRB time scales <1 ms, the dimension of the emission region had to be <300 km. This suggested compact objects such as neutron stars or black holes, or possibly relativistic shock waves, in close proximity to a source at similar scales. Hot extended plasmas emit incoherent radio emission because of the interaction of charged particles, with a spectrum and luminosity determined by their dimension and temperature. To reach the luminosities of FRBs in the available time scale would require an incoherent source to have an implausible temperature (~10^{40} K). Therefore, the FRB emission mechanism must be a coherent process, one in which N charged particles emit radio waves all in phase, producing N^2 times the power of a single particle (*48*). Examples of coherent processes are (i) plasma emission involving the generation of Langmuir waves and subsequent conversion into waves at the plasma frequency, (ii) electron cyclotron maser emission, and (iii) pulsars. Of these, the pulsar emission mechanism is the least understood. Giant pulses from the Crab Pulsar are thought to be caused by many nanosecond time-scale shots, each of which individually produces coherent emission that appear in quick succession, resulting in a giant pulse (*49*).

Early models for extragalactic FRBs could be assigned into two broad categories. In the first, some catastrophic explosion or other source-destroying event occurred, releasing a large amount of energy, some small fraction of which was converted into a coherent radio pulse. These are known as cataclysmic models. Examples include a neutron star–neutron star merger (*50*), a core-collapse gamma-ray burst or otherwise unusual (superluminous?) supernova explosion, or a neutron star that briefly exceeds its maximum stable mass before collapsing to a black hole (*51*). In these models, the FRB can

never repeat, and many of the events are expected to be associated with star-forming galaxies, which have large numbers of massive stars and high supernova rates. Cataclysmic models invoking decelerating blast waves (*52*) predicted unresolved FRBs with radio bandwidths $\delta\nu/\nu \sim 1$, similar to the bandwidths of the receivers used in many radio telescopes, whereas models invoking the magnetospheres of relativistic objects could contain finer temporal and band-limited spectral features (*49*), as exhibited in the giant pulses from energetic pulsars.

The second class of model was noncataclysmic, so it could allow FRBs to repeat. Giant pulses are the very bright (≳100 times the mean flux density) single pulses emitted by high-magnetic field young pulsars, including the Crab Pulsar (*53*) and PSR J0540–6919 (*54*), which rotate at ~20 to 30 Hz. Millisecond pulsars rotate at up to 700 Hz, and two examples that emit giant pulses are PSR B1937+21 (*55*) and PSR J1823–3021A, which is located in a globular cluster (*56*). FRBs are >1 billion times the luminosities of the most luminous giant pulses from the Crab Pulsar, although it has been suggested that FRBs could be a related phenomenon producing much rarer super-giant pulses (*49*) from highly energetic pulsars. The origin of the giant pulses is unclear, but if a pulsar's propensity to emit a giant pulse depends upon its magnetic field and spin period, then there could be extremely magnetic and rapidly spinning neutron stars in the Universe (called millisecond magnetars) that could emit numerous FRB-like pulses, albeit for a very short time (less than a year) before they exhaust their rotational kinetic energy. If this model is correct, then FRBs would not be one-off sources and would preferentially be located in the spiral arms of star-forming galaxies (like young pulsars), possibly inside supernova remnants. They might also exhibit an underlying quasi-periodicity, like giant pulses, which tend to appear at particular rotation phases. Magnetars (neutron stars with high magnetic fields) would be expected in star-forming regions, but millisecond pulsars are known to occur both within globular clusters and the disks and halos of galaxies (*57*), because their rotational kinetic energy is sufficient to power them for more than the age of the Universe.

Some models did not fit into either category, such as those implying that FRB sources are within the Milky Way with spurious *DM*, such as flare stare models (*58*, *59*). Although these models greatly reduced the required intrinsic luminosities, they required an alternative explanation for the dispersion of the pulses, necessitating a fine-tuned physical model to produce the observed dispersion sweep.

Discovery of repeating FRBs

Searches for repetition of the FRBs detected using Parkes found none after >100 hours of observation (*44*, *60*). However in 2016, FRB 121102, then the only FRB detected using Arecibo, was found to repeat, producing many repeat bursts in a single observing session (*61*). It was nicknamed the repeater (later known as R1).

In follow-up observations of FRB 121102, 10 additional bursts were discovered, two on 1 day, then eight on another, with six appearing during a 10-min period. On other days, no bursts were seen (*61*). The repeater's discovery ruled out the cataclysmic models, at least for for repeating FRBs. The bursts from FRB 121102 also appeared to come in clumps, unlike giant pulses from pulsars, which are more random (*56*, *62*).

FRBs from the repeater appeared to be subtly different than the nonrepeating FRBs observed with Parkes and the GBT. Bursts from the repeater often had multiple components and were broader in temporal extent (~5 versus ~1 ms). These components were often confined to small fractional bandwidths ($\delta\nu/\nu \sim 0.2$), with emission within each burst drifting to lower radio frequencies (Fig. 2B′). This behavior had been predicted for radio emission from magnetars (*63*) before any FRB had been discovered. The emission has been described as being like a sad trombone, with multiple notes each progressively lower in tone (*64*).

The repeater's active periods provided an opportunity for follow-up with interferometers to determine a precise location suitable for the identification of host galaxies. During one such active period, the realfast (*65*) instrument on the Very Large Array (VLA) interferometer detected an FRB (*66*) and localized it to near a persistent radio source and faint optical companion. Subsequent very long baseline interferometry further localized it to its host galaxy (*67*) and showed that it was situated coincident with the persistent radio source (*68*). The host was a tiny dwarf galaxy containing 40 million solar masses ($M_\odot$) in stars and gas, with a high specific star-formation rate of ~0.4 $M_\odot$ year^{-1}, with elemental abundances (metallicity) showing that it is still undergoing its first wave of star formation. The repeater's host galaxy was very similar to those of many long-duration gamma-ray bursts and superluminous supernovae (*69*). Both of those types of transient are associated with young massive stars of low metallicity. The redshift ($z = 0.193$) (*67*), and hence distance, of the repeater's host galaxy removed all remaining doubt that FRBs were at cosmological distances. FRB 121102 was at a distance of 972 Mpc, close to the maximum estimated for the Lorimer Burst (*1*).

The association with the persistent radio source raised several questions. Was the persistent source enabling the FRB emission, caused by the FRB, or merely coincident with it? Perhaps the persistent radio source was an accreting intermediate mass black hole or nearby supernova remnants powered by young neutron stars produced by recently exploded stars?

Follow-up observations of the repeater detected 93 further bursts (*70*, *71*) between 4 and 8 GHz. These observations showed that, in its rest frame, FRB emission could extend to almost 10 GHz, but often with small fractional bandwidths. The repeater has a rotation measure of ~10^5 rad m^{-2}, among the highest known for any astronomical source (*72*). This implied that it was in a highly magnetized environment, similar to the Galactic Center of the Milky Way, which contains a supermassive black hole.

Subsequent observations using broadband receivers (*73*) demonstrated that the repeater's emission is often limited in frequency extent, with emission confined to the same finite radio bands for extended periods, which must be instrinsic to the source. Monitoring of the rotation measure has shown it varies over years, dropping to two-thirds of its original value, indicating a rapidly evolving magnetic environment (*74*).

Questions posed by the discovery of the repeater were: Do all FRBs repeat if observed for long enough? Are there two classes of FRB, repeaters and nonrepeaters? If every FRB source emits millions of FRBs (or more), then the formation rate of the sources could be very low, so the sources could potentially be highly exotic objects that are unknown from observations of the local Universe.

2017–2022: The FRB age of discovery

Most of the early FRBs were found with standard radio pulsar search instrumentation. Once the cosmological population was established, plans were made to accelerate the discovery rate using large field of view instruments. Some FRBs, such as the Lorimer Burst, were so bright that they should have been detectable by small dishes. Although small telescopes have much lower sensitivity, they also have wider fields of view than large dishes and are less expensive to build and operate. Purpose-built FRB facilities, both large and small, were constructed.

The UTMOST upgrade (Fig. 2C) of the large cylindrical Molonglo Observatory Synthesis Telescope (MOST) allowed an interferometer to identify FRBs in a blind survey (*75*). The Australian Square Kilometre Array Pathfinder (ASKAP) array (Fig. 2D) added an incoherent fast sampling mode to the 30-square-degree field of view (provided by its phased array feeds) and began detecting FRBs routinely, finding 20 in a fly's eye survey (*76*). The phased array feeds removed the degeneracy between FRB flux and (usually unknown) position in the primary beam, enabling measurements of absolute flux densities for one-off FRBs. The nearby high-flux FRBs detected with ASKAP

were shown to be local versions of fainter FRBs detected by the more sensitive Parkes dish in the more distant Universe (*76*). The expected d log N/d log S = −3/2 flux density distribution was beginning to be consistent with a cosmological population.

Many of these purpose-built facilities had the ability to buffer and store the raw data when an FRB occurred, which was determined automatically in real time. This enabled microsecond-time-resolution studies of FRB profiles, which revealed detailed microstructure down to only a few tens of microseconds (*77*, *78*) (Figs. 2, C′ and D′). These short time scales indicate that the emission occurs on subkilometer scales, consistent with gaps in neutron star magnetospheres (*79*).

The Canadian Hydrogen Intensity Mapping Experiment (CHIME) telescope in Canada (Fig. 2E) has a very wide field of view (~200 square degrees) and high instantaneous sensitivity (collecting area of 8000 m^2). This is complemented by the ASKAP interferometer's sub-arc-second localizing capabilities and the even higher sensitivity provided by the 500-m Five-Hundred Meter Aperture Spherical Telescope (FAST) dish (Fig. 2F). A very different approach was taken by Survey for Transient Astronomical Radio Emission 2 (STARE2), which operates three 20-cm coaxial feeds (Fig. 2G), which have less than a millionth of FAST's collecting area but view the entire sky.

CHIME (*80*) is a fixed 4 ×100 m long × 20 m wide cylindrical interferometer that forms 1024 coherent beams over the sky. Its FRB subproject CHIME/FRB searches for dispersed radio pulses almost continuously from 400 to 800 MHz, scanning the entire northern sky every day. The CHIME/FRB average detection rate (a few per day) has rapidly increased the catalog of known FRBs, which now has >600 unique sources. CHIME observations have provided insights into FRB emission at low (400 to 800 MHz) radio frequencies (*81*) and identified a large number of repeaters (*82*–*84*), many of which were localized with other facilities.

In early 2020, CHIME detected two FRB-like bursts of emission (*85*) from the Galactic magnetar SGR 1935+2154 separated by only 30 ms. Just a second earlier, STARE2 also detected a millisecond-duration radio burst (*86*) (Fig. 2G′) at 1.4 GHz. The time delays between the two instruments were consistent with the delay caused by pulse dispersion. These radio bursts were coincident with an x-ray burst from the magnetar (*87*). Although the radio luminosity of the burst was 30 times weaker than the least luminous FRB then known, it demonstrated that magnetars could emit FRB-like emission. At least some FRBs, and possibly all, are emitted from magnetars.

The initial trillion-fold luminosity gap between the Lorimer Burst and Galactic pulsars has gradually closed, although a gap still persists (Fig. 3). The definition of what constitutes an FRB is also becoming blurred. Early observations of FRBs were often temporally smeared by the instrument and had approximately millisecond durations; however, as instrumentation improved, both intrinsically narrower and broader FRBs were observed. Periodic emission (*88*) has been reported in an unusually long ~3-s burst of radio emission, with periodic spikes at separations of 216.8 ms (Fig. 2E′). Two other FRBs have potential periodic emission with periods of 2.8 and 10.7 ms (*88*). Do these reflect the rotation periods of neutron star hosts, or are FRBs similar to some radio magnetars that often exhibit spiky emission (*89*)?

Repeaters galore

CHIME detected a second repeater, FRB 20180814A, which has a low *DM* of 189 pc cm^{-3}, indicating a distance of only 350 Mpc (*82*). The detections increased so rapidly that there were soon another 17 repeaters (*83*, *84*). In its first year of operation, CHIME detected a total of 18 repeaters (3.7%) and 474 FRBs that were not seen to repeat (96.3%) (*90*). The repeating FRBs have narrower fractional radio bandwidths and are wider than one-off FRBs (*91*). Follow-up interferometric observations (*92*) demonstrated that one of these repeaters (FRB 20180916B) was emitted from within a nearby spiral galaxy at a redshift of z = 0.0337 (~170 Mpc). This FRB was six times closer than the original repeater (R1) and less luminous that previously observed extragalactic FRBs.

Analysis of the repeater FRB 180916B showed the bursts were all received in a 5-day-wide window that recurred every 16.35 days, with more than half concentrated in a narrower 0.6-day-wide window (*93*). This motivated searches for periodicities in other repeaters. R1's bursts were then shown to be consistent with a 157 to 161 d periodicity with a broader fractional activity cycle (~50%) (*94*, *95*). A consistent periodicity usually indicates either two stars in an orbit or precession, the reorientation of a spin axis (tracing out a conical shape). This led to the suggestion that repeating FRBs were magnetars orbiting other active (massive?) stars in such a way so that they are only observable at certain orbital phases (*96*). This model received some support when the radio frequency of bursts from repeating FRBs was shown to be (orbital?) phase dependent, with the lower-frequency FRBs coming later in the cycle than the high-frequency ones (*97*, *98*). Others (*99*) have suggested that repeating FRBs might be induced by material streaming past magnetars. These types of models are illustrated in Fig. 4.

Observations with FAST (Fig. 2F) detected 1652 FRBs from the original repeater in <60 hours of observation time (*100*). During its most active observed period, the repeater was bursting on average every 30 s. No periodicity was found, and the high repetition rate means that the radio emission mechanism must be very efficient. The *DM* of the repeater is increasing, not decreasing, as would be expected if it were at the center of an expanding supernova remnant (*101*, *102*).

Another repeater was found with a *DM* of only 87 pc cm^{-3}, associated with the spiral galaxy Messier 81 (M81) at a distance of ~3.6 Mpc (*92*). Interferometry showed that it is almost certainly associated with a globular cluster in M81. Most of the dispersion for this burst arises from the foreground Milky Way and within M81, not the intergalactic medium. This FRB is only about 0.4% of the distance to the original repeater, of similar brightness, and thus about five orders of magnitude less luminous. Globular clusters do not contain young stars, and any magnetars formed in the first wave of star formation are expected to have become inactive long ago. If this repeating FRB is also produced by a magnetar, then it might have been formed recently, either by the collapse of an accreting white dwarf or by the merger of two neutron stars (*92*). The FRBs might alternatively arise from a millisecond pulsar, which are abundant in globular clusters and are known to emit giant pulses (*56*, *103*). Follow-up observations of this source found burst storms, in which it emits an FRB more than once per minute, but with no associated periodicity (*104*). Unlike some of the repeaters that show dispersion measure variations, this source has a stable *DM* (*104*), consistent with the expected environment within a globular cluster.

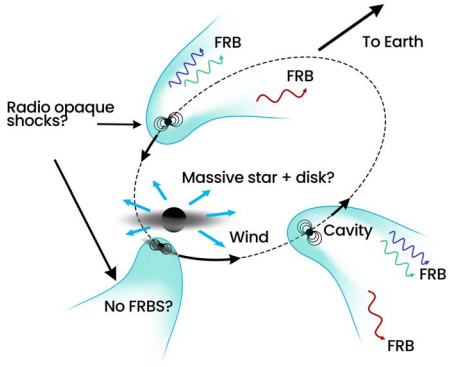


Fig. 4. Repeating FRB orbital model. In this repeating FRB model (*99*), a massive star's stellar wind (blue arrows) causes an orbiting (with a period of weeks to months) magnetar (small black dots) to emit FRBs. The interaction between the stellar wind and the magnetar wind produces cavities (cyan shading) separated by a shock front (cyan lines). The cavity preferentially emits high-frequency FRBs (blue and green wavy arrows) at the leading edge and lower-frequency FRBs (red wavy arrows) in the trailing sections. This model is an attempt to explain both the activity windows of some repeaters and their radio frequency time dependence.

A very high *DM* = 1205 pc cm^{-3} repeating FRB observed with FAST was subsequently localized to a galaxy (*105*), which is much closer than the intergalactic *DM* model suggested, implying that the host galaxy must be contributing ~75% of the total dispersion. Like the original repeater, this FRB is associated with a persistent radio source that is probably related to the anomalous *DM*. This source demonstrates the potential pitfalls of assuming that the *DM* reliably predicts distances.

The linear polarization fraction of repeating FRBs was shown to be strongly radio frequency dependent, as predicted if their radio waves are scattered in a highly variable magnetic environment (*106*). At lower frequencies, radio waves experience more variable Faraday rotation, leading to the observed systematic depolarization.

Host galaxies and cosmological applications

Advances in instrumentation have enabled interferometers to determine the precise locations of one-off FRBs, identifying their host galaxies (*107*–*109*), as well as following up repeaters. A study of six repeating and 10 nonrepeating FRBs with known host galaxies and redshifts (ranging from z = 0.008 to 0.66) found that there was no statistically significant difference between their host galaxies (*110*). However, the same study found that FRBs are rare in elliptical galaxies, being more common in galaxies that are experiencing at least some star formation (*110*). One-off FRBs are less common in galaxies with high star formation rates per unit mass, unlike long gamma-ray bursts, which are often associated with low-metallicity, low-mass hosts and produced by exploding massive stars. Nonrepeaters also appear to have different host galaxy properties to core-collapse supernovae (*110*). This is inconsistent with unification models proposing that all FRBs are from young magnetars produced in recent supernovae. Could nonrepeaters be produced by neutron stars reactivated long after their formation? One potential reactivation mechanism is mass transfer from a companion star. That process is known to produce millisecond pulsars, some of which emit giant pulses. Neutron stars in binaries accrete mass and gain angular momentum when their companions exhaust their fuel and swell up during stellar evolution. The length of the delay between neutron star birth and this accretion depends upon the mass of the companion star; it can be between 1 Myr and >1 Gyr. Millisecond pulsars could explain the host properties of one-off FRBs but not their lack of repetition.

Enough FRB host galaxies have been determined to derive an FRB redshift–*DM* relation (*111*) (also known as the Macquart relation). The observed relation is consistent with the total mass of baryons inferred by studies of the cosmic microwave background (*112*, *113*), possibly resolving the difficulty in locating all of the baryons using other methods (*29*).

The sheer numbers of FRBs detected by CHIME are providing other insights into the population. Analysis of the catalog of detected FRBs (*90*) found that the *DM* and flux density distributions are consistent with a cosmological population and with the large-scale structure of galaxies in the Universe (*114*).

FRB dispersion measures and host galaxy redshifts have been used to independently derive the value of the Hubble constant (H_0), the rate at which the expansion velocity of the Universe increases with distance. With a limited sample of nine FRBs with redshifts, a value of H_0 = 62 ± 9 km s^{-1} Mpc^{-1} has been deduced (*115*), albeit with some possibly optimistic assumptions about the (contaminating) host galaxy *DM* contributions. This is less precise than other methods; improvements will require eliminating FRBs with high local *DM* contributions, possibly by examining their intrinsic widths and scattering or by characterizing their host galaxy environments.

Current status and future prospects

So what are FRBs? My personal view is that, like many new classes of object, FRBs will ultimately be shown to be composed of one or two dominant sources, but there could be other rarer classes of source with the right combination of magnetic field, rotation, gravity, and accelerated charged particles to generate FRBs. Determining the locations of ~100 FRBs should provide sufficient information on their host galaxies to constrain the progenitors. My leading contenders for the repeaters are magnetars, with some in orbit around massive stars, whereas nonrepeaters seem more likely to be rare giant pulses from high-magnetic field (~10^9 G?) or recently spun-up millisecond pulsars. Both magnetars and millisecond pulsars experience magnetic field reconfigurations (Fig. 5) leading to changes in radio pulse shape changes. In magnetars, this can produce high-energy outbursts (*116*)—and in one case, a low-luminosity FRB (*85*, *86*). Could these magnetic reconfigurations be a common trigger for FRB production? In millisecond pulsars, these reconfigurations are extremely rare, and of the few hundred known millisecond pulsars, only a few have been seen to exhibit them, including PSR J1713+0747 (*117*). If this model is correct, then eventually all FRBs might repeat, but we might have to wait decades or more to observe them.

I expect that progress in the field will be strongly linked to new facilities coming online in the next decade. The MeerTRAP (*118*) experiment is predicted to detect and localize FRBs at higher distances. The CRACO (Commensal Real-time ASKAP Fast Transients Coherent) upgrade to the ASKAP interferometer (set to become operational in late 2022) will coherently add the signals from the inner 30 antennas to improve the FRB localization rate by

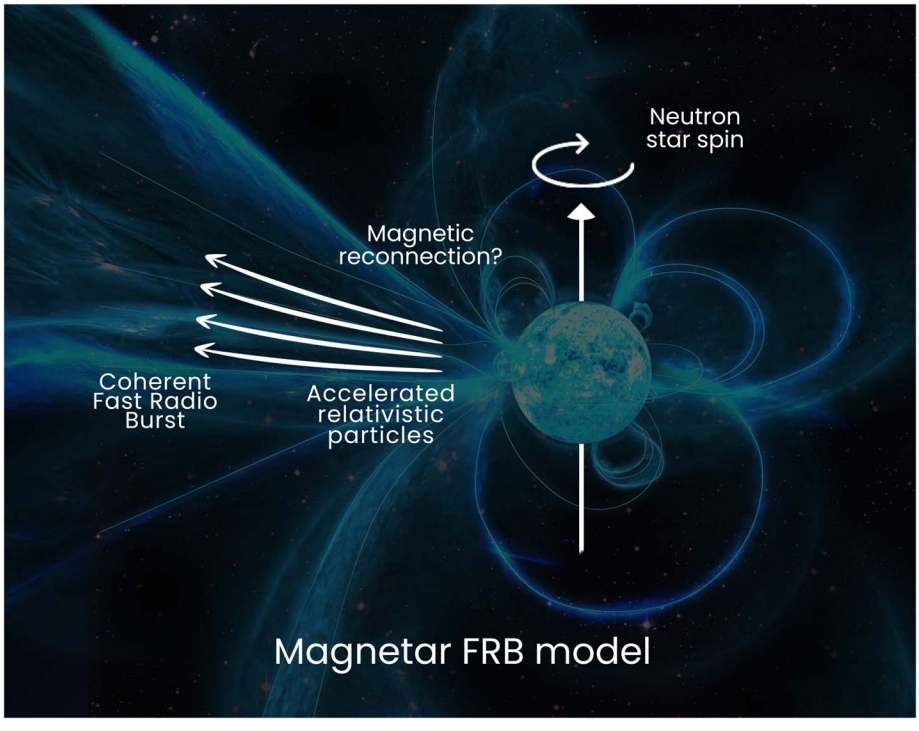


Fig. 5. Magnetar FRB emission model. Reconfiguration of the intense magnetic fields around a magnetar is associated with high-energy outbursts. In this model for FRB generation, reconfiguration of the magnetic field releases relativistic particles that generate coherent radio emission in the magnetosphere, possibly producing FRBs.

an order of magnitude, whereas the Deep Synoptic Array (DSA 110) (*119*) (also coming in late 2022) is predicted to localize almost one FRB per day. The CHIME/FRB outrigger project is deploying additional cylinders to enable localization. The Canadian Hydrogen Observatory and Radio-transient Detector (CHORD) (*120*) will be an array of 512 6-m dishes supported by outrigger stations that will enable rapid localization of FRBs and should be operational some time in the mid 2020s. Farther in the future, instruments including the DSA 2000 (*121*) (circa 2027) plan to localize hundreds of FRBs per day and the Square Kilometre Array (*122*) (from 2030) is expected to probe high-redshift FRBs. The results from these new instruments will ensure that the golden age of FRB discovery extends well into the 2030s.

Would FRBs have ever been discovered if not for the brightness of the Lorimer Burst? Parallels have been drawn with the first gravitational wave source, which remains the one with the largest known amplitude (*123*, *124*). The Lorimer Burst was only detected because it appeared in the sidelobes of the telescope, so in that sense, its high flux density was necessary for the discovery. However the high fluence of the burst also compromised both its detection and its acceptance. The simple interference rejection algorithm in operation tried to erase the burst, and its brightness led to arguments that it was statistically unlikely, and hence probably interference. I am convinced that the visual recognition of the Lorimer Burst's dispersed bright pulse played a role in convincing astronomers that FRBs existed, but I am also certain that it was the scientific potential of FRBs that motivated the instrumental developments and surveys that ultimately produced the scientific discoveries discussed in this Review.

REFERENCES AND NOTES

1. D. R. Lorimer, M. Bailes, M. A. McLaughlin, D. J. Narkevic, F. Crawford, A bright millisecond radio burst of extragalactic origin. *Science* **318**, 777–780 (2007). doi: 10.1126/science.1147532; pmid: 17901298
2. A. Hewish, S. J. Bell, J. D. H. Pilkington, P. F. Scott, R. A. Collins, Observation of a Rapidly Pulsating Radio Source. *Nature* **217**, 709–713 (1968). doi: 10.1038/217709a0
3. P. Goldreich, W. H. Julian, Pulsar Electrodynamics. *Astrophys. J.* **157**, 869 (1969). doi: 10.1086/150119
4. M. Ruderman, Pulsars: Structure and Dynamics. *Annu. Rev. Astron. Astrophys.* **10**, 427–476 (1972). doi: 10.1146/annurev.aa.10.090172.002235
5. D. R. Lorimer, M. Kramer, *Handbook of Pulsar Astronomy* (Cambridge Univ. Press) (2012).
6. R. N. Manchester *et al*., The Parkes multi-beam pulsar survey - I. Observing and data analysis systems, discovery and timing of 100 pulsars. *Mon. Not. R. Astron. Soc.* **328**, 17–35 (2001). doi: 10.1046/j.1365-8711.2001.04751.x
7. D. J. Morris *et al*., The Parkes Multibeam Pulsar Survey - II. Discovery and timing of 120 pulsars. *Mon. Not. R. Astron. Soc.* **335**, 275–290 (2002). doi: 10.1046/j.1365-8711.2002.05551.x
8. M. Kramer *et al*., The Parkes Multibeam Pulsar Survey. III. Young pulsars and the discovery and timing of 200 pulsars. *Mon. Not. R. Astron. Soc.* **342**, 1299–1324 (2003). doi: 10.1046/j.1365-8711.2003.06637.x
9. F. Crawford *et al*., Radio pulsars in the Magellanic clouds. *Astrophys. J.* **553**, 367–374 (2001). doi: 10.1086/320635
10. R. A. Hulse, J. H. Taylor, Discovery of a pulsar in a binary system. *Astrophys. J.* **195**, L51–L53 (1975). doi: 10.1086/181708
11. D. C. Backer, S. R. Kulkarni, C. Heiles, M. M. Davis, W. M. Goss, A millisecond pulsar. *Nature* **300**, 615–618 (1982). doi: 10.1038/300615a0
12. L. Staveley-Smith *et al*., The Parkes 21 CM multibeam receiver. *Publ. Astron. Soc. Aust.* **13**, 243–248 (1996). doi: 10.1017/S1323358000020919
13. R. T. Edwards, M. Bailes, W. van Straten, M. C. Britton, The Swinburne intermediate-latitude pulsar survey. *Mon. Not. R. Astron. Soc.* **326**, 358–374 (2001). doi: 10.1046/j.1365-8711.2001.04637.x
14. R. N. Manchester, G. Fan, A. G. Lyne, V. M. Kaspi, F. Crawford, Discovery of 14 radio pulsars in a survey of the magellanic clouds. *Astrophys. J.* **649**, 235–242 (2006). doi: 10.1086/505461
15. M. A. McLaughlin *et al*., Transient radio bursts from rotating neutron stars. *Nature* **439**, 817–820 (2006). doi: 10.1038/nature04440; pmid: 16482150
16. P. Weltevrede, B. W. Stappers, J. M. Rankin, G. A. E. Wright, Is Pulsar B0656+14 a very nearby rotating radio transient? *Astrophys. J.* **645**, L149–L152 (2006). doi: 10.1086/506346
17. E. F. Keane *et al*., Further searches for rotating radio transients in the Parkes Multi-beam Pulsar Survey. *Mon. Not. R. Astron. Soc.* **401**, 1057–1068 (2010). doi: 10.1111/j.1365-2966.2009.15693.x
18. S. Burke-Spolaor, M. Bailes, The millisecond radio sky: Transients from a blind single-pulse search. *Mon. Not. R. Astron. Soc.* **402**, 855–866 (2010). doi: 10.1111/j.1365-2966.2009.15965.x
19. S. R. Kulkarni, Dispersion measure: Confusion, constants & clarity. arXiv:2007.02886 [astro-ph.HE] (2020).
20. J. M. Cordes, T. J. W. Lazio, NE2001.I. A new model for the galactic distribution of free electrons and its fluctuations. arXiv:astro-ph/0207156 [astro-ph] (2002).
21. J. M. Yao, R. N. Manchester, N. Wang, A new electron-density model for estimation of pulsar and FRB distances. *Astrophys. J.* **835**, 29 (2017). doi: 10.3847/1538-4357/835/1/29
22. K. Ioka, The cosmic dispersion measure from gamma-ray burst afterglows: Probing the reionization history and the burst environment. *Astrophys. J.* **598**, L79–L82 (2003). doi: 10.1086/380598
23. A. W. Hotan, W. van Straten, R. N. Manchester, PSRCHIVE and PSRFITS: An open approach to radio pulsar data storage and analysis. *Publ. Astron. Soc. Aust.* **21**, 302–309 (2004). doi: 10.1071/AS04022
24. W. van Straten, M. Bailes, DSPSR: Digital signal processing software for pulsar astronomy. *Publ. Astron. Soc. Aust.* **28**, 1–14 (2011). doi: 10.1071/AS10021
25. N. D. R. Bhat, J. M. Cordes, F. Camilo, D. J. Nice, D. R. Lorimer, Multifrequency observations of radio pulse broadening and constraints on interstellar electron density microstructure. *Astrophys. J.* **605**, 759–783 (2004). doi: 10.1086/382680
26. E. Petroff *et al*., FRBCAT: The fast radio burst catalogue. *Publ. Astron. Soc. Aust.* **33**, e045 (2016). doi: 10.1017/pasa.2016.35
27. V. Ravi, The observed properties of fast radio bursts. *Mon. Not. R. Astron. Soc.* **482**, 1966–1978 (2019). doi: 10.1093/mnras/sty1551
28. S. B. Zhang *et al*., A new fast radio burst in the data sets containing the Lorimer burst. *Mon. Not. R. Astron. Soc.* **484**, L147–L150 (2019). doi: 10.1093/mnrasl/slz023
29. J. M. Shull, B. D. Smith, C. W. Danforth, The Baryon Census in a multiphase intergalactic medium: 30% of the Baryons may still be missing. *Astrophys. J.* **759**, 23 (2012). doi: 10.1088/0004-637X/759/1/23
30. M. Bagchi, A. C. Nieves, M. McLaughlin, A search for dispersed radio bursts in archival Parkes Multibeam Pulsar Survey data. *Mon. Not. R. Astron. Soc.* **425**, 2501–2506 (2012). doi: 10.1111/j.1365-2966.2012.21708.x
31. A. P. V. Siemion *et al*., The Allen Telescope Array Fly's Eye Survey for Fast Radio Transients. *Astrophys. J.* **744**, 109 (2012). doi: 10.1088/0004-637X/744/2/109
32. R. S. Lynch *et al*., The Green Bank Telescope 350 MHz Drift-scan Survey II: Data Analysis and the timing of 10 new pulsars, including a relativistic binary. *Astrophys. J.* **763**, 81 (2013). doi: 10.1088/0004-637X/763/2/81
33. S. Burke-Spolaor, M. Bailes, R. Ekers, J.-P. Macquart, I. CrawfordIII, Fronefield, Radio bursts with extragalactic spectral characteristics show terrestrial origins. *Astrophys. J.* **727**, 18 (2011). doi: 10.1088/0004-637X/727/1/18
34. E. Petroff *et al*., Identifying the source of perytons at the Parkes radio telescope. *Mon. Not. R. Astron. Soc.* **451**, 3933–3940 (2015). doi: 10.1093/mnras/stv1242
35. E. F. Keane, B. W. Stappers, M. Kramer, A. G. Lyne, On the origin of a highly dispersed coherent radio burst. *Mon. Not. R. Astron. Soc.* **425**, L71–L75 (2012). doi: 10.1111/j.1745-3933.2012.01306.x
36. K. W. Bannister, G. J. Madsen, A Galactic origin for the fast radio burst FRB010621. *Mon. Not. R. Astron. Soc.* **440**, 353–358 (2014). doi: 10.1093/mnras/stu220
37. M. J. Keith *et al*., The High Time Resolution Universe Pulsar Survey. I. System configuration and initial discoveries. *Mon. Not. R. Astron. Soc.* **409**, 619–627 (2010). doi: 10.1111/j.1365-2966.2010.17325.x
38. J. Kocz, M. Bailes, D. Barnes, S. Burke-Spolaor, L. Levin, Enhanced pulsar and single pulse detection via automated radio frequency interference detection in multipixel feeds. *Mon. Not. R. Astron. Soc.* **420**, 271–278 (2012). doi: 10.1111/j.1365-2966.2011.20029.x
39. D. Thornton *et al*., A population of fast radio bursts at cosmological distances. *Science* **341**, 53–56 (2013). doi: 10.1126/science.1236789; pmid: 23828936
40. L. G. Spitler *et al*., Fast radio burst discovered in the Arecibo Pulsar ALFA Survey. *Astrophys. J.* **790**, 101 (2014). doi: 10.1088/0004-637X/790/2/101
41. S. Burke-Spolaor, K. W. Bannister, The galactic position dependence of fast radio bursts and the discovery of FRB011025. *Astrophys. J.* **792**, 19 (2014). doi: 10.1088/0004-637X/792/1/19
42. K. Masui *et al*., Dense magnetized plasma associated with a fast radio burst. *Nature* **528**, 523–525 (2015). doi: 10.1038/nature15769; pmid: 26633633
43. D. J. Champion *et al*., Five new fast radio bursts from the HTRU high-latitude survey at Parkes: First evidence for two-component bursts. *Mon. Not. R. Astron. Soc.* **460**, L30–L34 (2016). doi: 10.1093/mnrasl/slw069
44. V. Ravi, R. M. Shannon, A. Jameson, A fast radio burst in the direction of the Carina Dwarf Spheroidal Galaxy. *Astrophys. J. Lett.* **799**, L5 (2015). doi: 10.1088/2041-8205/799/1/L5
45. V. Ravi *et al*., The magnetic field and turbulence of the cosmic web measured using a brilliant fast radio burst. *Science* **354**, 1249–1252 (2016). doi: 10.1126/science.aaf6807; pmid: 27856844
46. S. R. Kulkarni, E. O. Ofek, J. D. Neill, Z. Zheng, M. Juric, Giant sparks at cosmological distances? *Astrophys. J.* **797**, 70 (2014). doi: 10.1088/0004-637X/797/1/70
47. E. Platts *et al*., A living theory catalogue for fast radio bursts. *Phys. Rep.* **821**, 1–27 (2019). doi: 10.1016/j.physrep.2019.06.003
48. D. B. Melrose, Coherent emission mechanisms in astrophysical plasmas. *Rev. Mod. Plasma Phys.* **1**, 5 (2017). doi: 10.1007/s41614-017-0007-0
49. J. M. Cordes, I. Wasserman, Supergiant pulses from extragalactic neutron stars. *Mon. Not. R. Astron. Soc.* **457**, 232–257 (2016). doi: 10.1093/mnras/stv2948
50. T. Totani, Cosmological fast radio bursts from binary neutron star mergers. *Publ. Astron. Soc. Jpn.* **65**, L12 (2013). doi: 10.1093/pasj/65.5.L12
51. H. Falcke, L. Rezzolla, Fast radio bursts: The last sign of supramassive neutron stars. *Astron. Astrophys.* **562**, A137 (2014). doi: 10.1051/0004-6361/201321996
52. B. D. Metzger, B. Margalit, L. Sironi, Fast radio bursts as synchrotron maser emission from decelerating relativistic blast waves. *Mon. Not. R. Astron. Soc.* **485**, 4091–4106 (2019). doi: 10.1093/mnras/stz700
53. T. H. Hankins, J. S. Kern, J. C. Weatherall, J. A. Eilek, Nanosecond radio bursts from strong plasma turbulence in the Crab pulsar. *Nature* **422**, 141–143 (2003). doi: 10.1038/nature01477; pmid: 12634779
54. M. Geyer *et al*., The Thousand-Pulsar-Array programme on MeerKAT - III. Giant pulse characteristics of PSR J0540-6919. *Mon. Not. R. Astron. Soc.* **505**, 4468–4482 (2021). doi: 10.1093/mnras/stab1501
55. I. Cognard, J. A. Shrauner, J. H. Taylor, S. E. Thorsett, Giant radio pulses from a millisecond pulsar. *Astrophys. J.* **457**, L81 (1996). doi: 10.1086/309894
56. F. Abbate *et al*., Giant pulses from J1823-3021A observed with the MeerKAT telescope. *Mon. Not. R. Astron. Soc.* **498**, 875–882 (2020). doi: 10.1093/mnras/staa2510
57. R. T. Bartels, T. D. P. Edwards, C. Weniger, Bayesian model comparison and analysis of the Galactic disc population of gamma-ray millisecond pulsars. *Mon. Not. R. Astron. Soc.* **481**, 3966–3987 (2018). doi: 10.1093/mnras/sty2529
58. A. Loeb, Y. Shvartzvald, D. Maoz, Fast radio bursts may originate from nearby flaring stars. *Mon. Not. R. Astron. Soc. Lett.* **439**, L46–L50 (2014). doi: 10.1093/mnrasl/slt177

59. B. Dennison, Fast radio bursts: Constraints on the dispersing medium. *Mon. Not. R. Astron. Soc. Lett.* **443**, L11–L14 (2014). doi: 10.1093/mnrasl/slu072
60. E. Petroff *et al.*, A survey of FRB fields: Limits on repeatability. *Mon. Not. R. Astron. Soc.* **454**, 457–462 (2015). doi: 10.1093/mnras/stv1953
61. L. G. Spitler *et al.*, A repeating fast radio burst. *Nature* **531**, 202–205 (2016). doi: 10.1038/nature17168; pmid: 26934226
62. S. C. Lundgren *et al.*, Giant pulses from the Crab Pulsar: A joint radio and gamma-ray study. *Astrophys. J.* **453**, 433 (1995). doi: 10.1086/176404
63. M. Lyutikov, Radio emission from magnetars. *Astrophys. J.* **580**, L65–L68 (2002). doi: 10.1086/345493
64. F. Rajabi, M. A. Chamma, C. M. Wyenberg, A. Mathews, M. Houde, A simple relationship for the spectro-temporal structure of bursts from FRB 121102. *Mon. Not. R. Astron. Soc.* **498**, 4936–4942 (2020). doi: 10.1093/mnras/staa2723
65. C. J. Law *et al.*, realfast: Real-time, commensal fast transient surveys with the very large array. *Astrophys. J. Suppl. Ser.* **236**, 8 (2018). doi: 10.3847/1538-4365/aab77b
66. S. Chatterjee *et al.*, A direct localization of a fast radio burst and its host. *Nature* **541**, 58–61 (2017). doi: 10.1038/nature20797; pmid: 28054614
67. S. P. Tendulkar *et al.*, The host galaxy and redshift of the repeating fast radio burst FRB 121102. *Astrophys. J. Lett.* **834**, L7 (2017). doi: 10.3847/2041-8213/834/2/L7
68. B. Marcote *et al.*, The repeating fast radio burst FRB 121102 as seen on milliarcsecond angular scales. *Astrophys. J. Lett.* **834**, L8 (2017). doi: 10.3847/2041-8213/834/2/L8
69. Y. Li, B. Zhang, A comparative study of host galaxy properties between fast radio bursts and stellar transients. *Astrophys. J. Lett.* **899**, L6 (2020). doi: 10.3847/2041-8213/aba907
70. V. Gajjar *et al.*, Highest frequency detection of FRB 121102 at 4-8 GHz using the breakthrough listen digital backend at the Green Bank Telescope. *Astrophys. J.* **863**, 2 (2018). doi: 10.3847/1538-4357/aad005
71. Y. G. Zhang *et al.*, Fast Radio Burst 121102 pulse detection and periodicity: A machine learning approach. *Astrophys. J.* **866**, 149 (2018). doi: 10.3847/1538-4357/aadf31
72. D. Michilli *et al.*, An extreme magneto-ionic environment associated with the fast radio burst source FRB 121102. *Nature* **553**, 182–185 (2018). doi: 10.1038/nature25149; pmid: 29323297
73. K. Gourdji *et al.*, A sample of low-energy bursts from FRB 121102. *Astrophys. J. Lett.* **877**, L19 (2019). doi: 10.3847/2041-8213/ab1f8a
74. G. H. Hilmarsson *et al.*, Rotation measure evolution of the repeating fast radio burst source FRB 121102. *Astrophys. J. Lett.* **908**, L10 (2021). doi: 10.3847/2041-8213/abdec0
75. M. Caleb *et al.*, The first interferometric detections of fast radio bursts. *Mon. Not. R. Astron. Soc.* **468**, 3746–3756 (2017). doi: 10.1093/mnras/stx638
76. R. M. Shannon *et al.*, The dispersion-brightness relation for fast radio bursts from a wide-field survey. *Nature* **562**, 386–390 (2018). doi: 10.1038/s41586-018-0588-y; pmid: 30305732
77. W. Farah *et al.*, FRB microstructure revealed by the real-time detection of FRB170827. *Mon. Not. R. Astron. Soc.* **478**, 1209–1217 (2018). doi: 10.1093/mnras/sty1122
78. H. Cho *et al.*, Spectropolarimetric analysis of FRB 181112 at microsecond resolution: Implications for fast radio burst emission mechanism. *Astrophys. J. Lett.* **891**, L38 (2020). doi: 10.3847/2041-8213/ab7824
79. K. S. Cheng, C. Ho, M. Ruderman, Energetic radiation from rapidly spinning pulsars. I. Outer magnetosphere gaps. *Astrophys. J.* **300**, 500 (1986). doi: 10.1086/163829
80. M. Amiri *et al.*, The CHIME Fast Radio Burst Project: System overview. *Astrophys. J.* **863**, 48 (2018). doi: 10.3847/1538-4357/aad188
81. The CHIME/FRB Collaboration, Observations of fast radio bursts at frequencies down to 400 megahertz. *Nature* **566**, 230–234 (2019). doi: 10.1038/s41586-018-0867-7; pmid: 30653191
82. The CHIME/FRB Collaboration, A second source of repeating fast radio bursts. *Nature* **566**, 235–238 (2019). doi: 10.1038/s41586-018-0864-x; pmid: 30653190
83. The CHIME/FRB Collaboration *et al.*, CHIME/FRB discovery of eight new repeating fast radio burst sources. *Astrophys. J. Lett.* **885**, L24 (2019). doi: 10.3847/2041-8213/ab4a80
84. E. Fonseca *et al.*, Nine new repeating fast radio burst sources from CHIME/FRB. *Astrophys. J. Lett.* **891**, L6 (2020). doi: 10.3847/2041-8213/ab7208
85. The CHIME/FRB Collaboration, A bright millisecond-duration radio burst from a Galactic magnetar. *Nature* **587**, 54–58 (2020). doi: 10.1038/s41586-020-2863-y; pmid: 33149292
86. C. D. Bochenek *et al.*, A fast radio burst associated with a galactic magnetar. *Nature* **587**, 59–62 (2020). doi: 10.1038/s41586-020-2872-x; pmid: 33149288
87. D. M. Palmer, B. A. T. Team, A forest of bursts from SGR 1935+2154. *GRB Coordinates Network* **27665**, 1 (2020).
88. B. C. Andersen *et al.*, Sub-second periodicity in a fast radio burst. *Nature* **607**, 256–259 (2022). doi: 10.1038/s41586-022-04841-8; pmid: 35831603
89. L. Levin *et al.*, Radio emission evolution, polarimetry and multifrequency single pulse analysis of the radio magnetar PSR J1622-4950. *Mon. Not. R. Astron. Soc.* **422**, 2489–2500 (2012). doi: 10.1111/j.1365-2966.2012.20807.x
90. M. Amiri *et al.*, The first CHIME/FRB fast radio burst catalog. *Astrophys. J. Suppl. Ser.* **257**, 59 (2021). doi: 10.3847/1538-4365/ac33ab
91. Z. Pleunis *et al.*, Fast radio burst morphology in the first CHIME/FRB catalog. *Astrophys. J.* **923**, 1 (2021). doi: 10.3847/1538-4357/ac33ac
92. F. Kirsten *et al.*, A repeating fast radio burst source in a globular cluster. *Nature* **602**, 585–589 (2022). doi: 10.1038/s41586-021-04354-w; pmid: 35197615
93. The CHIME/FRB Collaboration, Periodic activity from a fast radio burst source. *Nature* **582**, 351–355 (2020). doi: 10.1038/s41586-020-2398-2; pmid: 32555491
94. K. M. Rajwade *et al.*, Possible periodic activity in the repeating FRB 121102. *Mon. Not. R. Astron. Soc.* **495**, 3551–3558 (2020). doi: 10.1093/mnras/staa1237
95. M. Cruces *et al.*, Repeating behaviour of FRB 121102: Periodicity, waiting times, and energy distribution. *Mon. Not. R. Astron. Soc.* **500**, 448–463 (2021). doi: 10.1093/mnras/staa3223
96. M. Lyutikov, M. V. Barkov, D. Giannios, FRB periodicity: Mild pulsars in tight O/B-star binaries. *Astrophys. J. Lett.* **893**, L39 (2020). doi: 10.3847/2041-8213/ab87a4
97. Z. Pleunis *et al.*, LOFAR detection of 110-188 MHz emission and frequency-dependent activity from FRB 20180916B. *Astrophys. J. Lett.* **911**, L3 (2021). doi: 10.3847/2041-8213/abec72
98. I. Pastor-Marazuela *et al.*, Chromatic periodic activity down to 120 megahertz in a fast radio burst. *Nature* **596**, 505–508 (2021). doi: 10.1038/s41586-021-03724-8; pmid: 34433943
99. B. Zhang, A "cosmic comb" model of fast radio bursts. *Astrophys. J. Lett.* **836**, L32 (2017). doi: 10.3847/2041-8213/aa5ded
100. D. Li *et al.*, A bimodal burst energy distribution of a repeating fast radio burst source. *Nature* **598**, 267–271 (2021). doi: 10.1038/s41586-021-03878-5; pmid: 34645999
101. B. Margalit, B. D. Metzger, A concordance picture of FRB 121102 as a flaring magnetar embedded in a magnetized ion-electron wind nebula. *Astrophys. J. Lett.* **868**, L4 (2018). doi: 10.3847/2041-8213/aaedad
102. A. L. Piro, B. M. Gaensler, The dispersion and rotation measure of supernova remnants and magnetized stellar winds: Application to fast radio bursts. *Astrophys. J.* **861**, 150 (2018). doi: 10.3847/1538-4357/aac9bc
103. R. W. Romani, S. Johnston, Giant pulses from the millisecond pulsar B1821-24. *Astrophys. J.* **557**, L93–L96 (2001). doi: 10.1086/323415
104. K. Nimmo *et al.*, Burst timescales and luminosities as links between young pulsars and fast radio bursts. *Nat. Astron.* **6**, 393–401 (2022). doi: 10.1038/s41550-021-01569-9
105. C. H. Niu *et al.*, A repeating fast radio burst associated with a persistent radio source. *Nature* **606**, 873–877 (2022). doi: 10.1038/s41586-022-04755-5; pmid: 35676486
106. Y. Feng *et al.*, Frequency-dependent polarization of repeating fast radio bursts-implications for their origin. *Science* **375**, 1266–1270 (2022). doi: 10.1126/science.abl7759; pmid: 35298266
107. K. W. Bannister *et al.*, A single fast radio burst localized to a massive galaxy at cosmological distance. *Science* **365**, 565–570 (2019). doi: 10.1126/science.aaw5903; pmid: 31249136
108. V. Ravi *et al.*, A fast radio burst localized to a massive galaxy. *Nature* **572**, 352–354 (2019). doi: 10.1038/s41586-019-1389-7; pmid: 31266051
109. C. J. Law *et al.*, A distant fast radio burst associated with its host galaxy by the very large array. *Astrophys. J.* **899**, 161 (2020). doi: 10.3847/1538-4357/aba4ac
110. S. Bhandari *et al.*, Characterizing the fast radio burst host galaxy population and its connection to transients in the local and extragalactic universe. *Astron. J.* **163**, 69 (2022). doi: 10.3847/1538-3881/ac3aec
111. J. P. Macquart *et al.*, A census of baryons in the Universe from localized fast radio bursts. *Nature* **581**, 391–395 (2020). doi: 10.1038/s41586-020-2300-2; pmid: 32461651
112. R. J. Cooke, M. Pettini, C. C. Steidel, One percent determination of the primordial deuterium abundance. *Astrophys. J.* **855**, 102 (2018). doi: 10.3847/1538-4357/aaab53
113. N. Aghanim *et al.*, Planck 2018 results. VI. Cosmological parameters. *Astron. Astrophys.* **641**, A6 (2020). doi: 10.1051/0004-6361/201833910
114. M. Rafiei-Ravandi *et al.*, CHIME/FRB catalog 1 results: Statistical cross-correlations with large-scale structure. *Astrophys. J.* **922**, 42 (2021). doi: 10.3847/1538-4357/ac1dab
115. S. Hagstotz, R. Reischke, R. Lilow, A new measurement of the Hubble constant using fast radio bursts. *Mon. Not. R. Astron. Soc.* **511**, 662–667 (2022). doi: 10.1093/mnras/stac077
116. F. Coti Zelati, N. Rea, J. A. Pons, S. Campana, P. Esposito, Systematic study of magnetar outbursts. *Mon. Not. R. Astron. Soc.* **474**, 961–1017 (2018). doi: 10.1093/mnras/stx2679
117. H. Xu *et al.*, A sustained pulse shape change in PSR J1713 +0747 possibly associated with timing and DM events. *The Astronomer's Telegram* **14642**, 1 (2021).
118. B. Stappers, "MeerTRAP: Real time commensal searching for transients and pulsars with MeerKAT," Paper presented at MeerKAT Science: On the Pathway to the SKA, Stellenbosch, South Africa, 25–27 May 2016.
119. J. Kocz *et al.*, DSA-10: A prototype array for localizing fast radio bursts. *Mon. Not. R. Astron. Soc.* **489**, 919–927 (2019). doi: 10.1093/mnras/stz2219
120. K. Vanderlinde *et al.*, "The Canadian Hydrogen Observatory and Radio-transient Detector (CHORD)," in *Canadian Long Range Plan for Astronomy and Astrophysics White Papers*, vol. 2020 (Zenodo, 2019); https://zenodo.org/record/3765414#.Y1GlGnbMKUk.
121. G. Hallinan *et al.*, The DSA-2000: A radio survey camera. arXiv:1907.07648 [astro-ph.IM] (2019).
122. T. Hashimoto *et al.*, Fast radio bursts to be detected with the square kilometre array. *Mon. Not. R. Astron. Soc.* **497**, 4107–4116 (2020). doi: 10.1093/mnras/staa2238
123. LIGO Scientific Collaboration, Virgo Collaboration, Observation of gravitational waves from a binary black hole merger. *Phys. Rev. Lett.* **116**, 061102 (2016). doi: 10.1103/PhysRevLett.116.061102; pmid: 26918975
124. LIGO Scientific Collaboration, Virgo Collaboration, KAGRA Collaboration, GWTC-3: Compact binary coalescences observed by LIGO and Virgo during the second part of the third observing run. arXiv:2111.03606 [gr-qc] (2021).
125. M. Bailes *et al.*, The MeerKAT telescope as a pulsar facility: System verification and early science results from MeerTime. *Publ. Astron. Soc. Aust.* **37**, e028 (2020). doi: 10.1017/pasa.2020.19
126. M. Bailes *et al.*, The UTMOST: A hybrid digital signal processor transforms the Molonglo Observatory Synthesis Telescope. *Publ. Astron. Soc. Aust.* **34**, e045 (2017). doi: 10.1017/pasa.2017.39
127. W. Zhu *et al.*, A fast radio burst discovered in FAST Drift Scan Survey. *Astrophys. J. Lett.* **895**, L6 (2020). doi: 10.3847/2041-8213/ab8e46
128. M. Caleb, E. Keane, A decade and a half of fast radio burst observations. *Universe* **7**, 453 (2021). doi: 10.3390/universe7110453

ACKNOWLEDGMENTS

I thank D. Lorimer for involving me in the pursuit of the Lorimer Burst; my colleagues in the HTRU, SUPERB, UTMOST, and CRAFT FRB collaborations; M. Caleb and C. Knox for help with the figures; A. Deller, K. Gourdji, C. Flynn, and R. Shannon for feedback on the manuscript; and S. Ransom and an anonymous referee for providing many important suggestions that greatly improved the manuscript. **Funding:** This work was supported by the Australian Research Council (Laureate Fellowship FL150100148) and the ARC Centre of Excellence for Gravitational Wave Discovery (OzGrav grant CE170100004). **Competing interests:** I declare no competing interests. **License information:** Copyright © 2022 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

Submitted 12 April 2022; accepted 5 October 2022
10.1126/science.abj3043

REPORT

ASTROPHYSICS

A limit on variations in the fine-structure constant from spectra of nearby Sun-like stars

Michael T. Murphy[1]*, Daniel A. Berke[1], Fan Liu (刘凡)[1], Chris Flynn[1,2], Christian Lehmann[1], Vladimir A. Dzuba[3], Victor V. Flambaum[3]

The fine structure constant α sets the strength of the electromagnetic force. The Standard Model of particle physics provides no explanation for its value, which could potentially vary. The wavelengths of stellar absorption lines depend on α but are subject to systematic effects owing to astrophysical processes in stellar atmospheres. We measured precise line wavelengths from observations of 17 stars, selected to have almost identical atmospheric properties to those of the Sun (solar twins), which reduces those systematic effects. We found that α varies by ≤50 parts per billion within 50 parsecs from Earth. Combining the results from all 17 stars provides an empirical local reference for stellar measurements of α, with an ensemble precision of 12 parts per billion.

The Standard Model of particle physics contains parameters known as fundamental constants. These include the coupling strengths of the known physical forces; the strength of electromagnetism is set by the fine-structure constant, $\alpha \equiv e^2/\hbar c$, where e is the elementary charge, $\hbar$ is the reduced Planck constant, and c is the (vacuum) speed of light. These dimensionless numbers are referred to as fundamental because the Standard Model does not predict their values. They are usually assumed to be universal constants—they do not depend on other (unknown) physics. Their values can only be established experimentally, and testing their constancy requires measurements under a wide range of physical conditions, such as different times, distances, and gravitational potentials. Measurements of laboratory atomic clocks have set an upper limit on relative variations in α to $\lesssim 10^{-18}$ year^{-1} over several years (*1*). On cosmological time and distance scales, absorption lines of distant gas clouds in the spectra of background quasars limit relative variations in α to $\lesssim$1 parts per million (ppm) (*2*–*4*). A study of giant stars within the Milky Way has set similar limits of $\lesssim$2 to 6 ppm (*5*, *6*).

Any variation in α would alter the energy levels of atoms and ions in characteristic ways (*7*). The rest-frame wave number of an absorption or emission line (ω_{obs}) would shift from its laboratory value (ω_0) in proportion to the relative change $\Delta\alpha/\alpha \equiv (\alpha_{obs} - \alpha_0)/\alpha_0$:

$$\frac{\Delta v}{c} \equiv \frac{\omega_0 - \omega_{obs}}{\omega_0} \approx -2\frac{\Delta\alpha}{\alpha}Q \qquad (1)$$

where α_0 and α_{obs} are the laboratory and observed values of α, respectively; Δv is the line shift in velocity units; and the sensitivity coefficient Q describes how much a given line shifts to the blue (for positive Q) or red. The approximation is valid for $\Delta\alpha/\alpha \ll 1$. In practice, the velocity shifts are measured for multiple lines and different atoms and ions, which is known as the many multiplet method. Using lines with a wide variety of Q coefficients increases the sensitivity to variations in α.

Sun-like stars are potentially suitable targets for the many multiplet method: Their spectra contain thousands of narrow, well-defined, strong (but unsaturated) absorption lines (Fig. 1A). The observed wavelengths of these lines could in principle be compared with their laboratory values while simultaneously accounting for the star's radial velocity. However, this simple approach is limited by large systematic errors; several physical mechanisms can shift the lines by up to ~700 m s^{-1} from their laboratory wavelengths, and the line profiles are asymmetric because they arise over a range of depths in stellar atmospheres (*8*, *9*). These effects produce velocity shifts (Δv) between lines, typically $\Delta v \sim 250$ m s^{-1} (*8*), which is equivalent to $\Delta\alpha/\alpha \sim 6$ ppm for a typical range in Q coefficients of ≈0.07 (*10*). Direct comparison of absorption lines in a single giant star to laboratory values has already reached this systematic error limit (*5*).

We adopted an alternative technique that compares absorption lines between stars that have intrinsically similar spectra, eliminating the need to compare with laboratory wavelengths. The atmospheric spectrum of an isolated main-sequence star depends primarily on its mass and heavy-element content, which determine three primary observable parameters: the effective temperature T_{eff}, iron metallicity [Fe/H], and surface gravity log g. We restrict our analysis to solar twins, which are defined as stars with these parameters within 100 K, 0.1 decimal exponent (dex), and 0.2 dex of the Sun's values, respectively. Spectra of two solar twins used in our analysis are shown in Fig. 1A. We measured the velocity-space separations of pairs of lines and then compared the same sets of lines between stars (Fig. 1B). This approach reduces the systematic errors from astrophysical line shifts and asymmetries because of the similarity of their stellar parameters. The use of pairs of lines removes any dependence on the stars' radial velocities, including any variations that could be caused by an orbiting companion (such as in a planetary or binary stellar system). For main-sequence stars, line shifts and asymmetries were observed to be correlated with the line's optical depth and wavelength (*8*), so we selected pairs with similar absorption depths (within 20%) and small separations (<800 km s^{-1}, equivalent to ≈13 Å at 5000 Å) (*11*–*13*). We chose these values to reduce the systematic effects while maintaining sensitivity to variations in α between stars.

We applied this solar twins method to archival solar twin spectra from the High Accuracy Radial velocity Planet Searcher (HARPS) spectrograph mounted on the European Southern Observatory (ESO) 3.6-m telescope at La Silla Observatory, Chile. HARPS is highly stable over time (*14*), and its wavelength scale has been precisely characterized by using laser frequency combs (*15*, *16*). This sets an instrumental systematic error limit of ~2 to 3 m s^{-1} in the velocity separations of line pairs (*12*). To reach this level, we restricted our analysis to HARPS exposures corrected for nonuniform detector pixel sizes (corrections, ~25 m s^{-1}) (*13*, *17*) and applied a further correction for sparsely sampled wavelength calibration (corrections, ~5 m s^{-1}) (*13*, *16*).

We selected 16 bright (nearby) solar twins with HARPS spectra, with signal-to-noise ratio (SNR) > 200 per 0.8 km s^{-1} pixel, plus the Sun through reflection of its light from the asteroid Vesta (with SNR > 150) (table S1) (*13*). With these SNRs, the statistical uncertainty in the velocity separation of two unresolved absorption lines is ~25 m s^{-1} from a single exposure, assuming that they absorb 50% of the stellar flux at their cores (*18*). The median number of HARPS exposures available was 10 exposures per star (range of 1 to 138), so by combining results from multiple exposures, we expected median statistical uncertainties to reduce to ~8 m s^{-1} per line pair, per star. By averaging over the sample of 17 stars, the uncertainty approaches that imposed by the available instrument calibration (*12*).

From 8843 lines listed in a solar atlas (*19*), we selected 22 that are separated from each other and not blended with other nearby

[1]Centre for Astrophysics and Supercomputing, Swinburne University of Technology, Hawthorn, Victoria 3122, Australia. [2]ARC Centre of Excellence for Gravitational Wave Discovery, Hawthorn, Victoria 3122, Australia. [3]School of Physics, University of New South Wales, Sydney, NSW 2052, Australia.
*Corresponding author. Email: mmurphy@swin.edu.au

stellar or telluric (Earth atmosphere) lines (*12*, *13*). All 22 lines are strong but unsaturated, absorbing 15 to 90% of the continuum in the HARPS spectrum of the Sun. The 22 lines, which form 17 different pairs (some share common lines), arise from the neutral atoms sodium (Na), calcium (Ca), titanium (Ti), vanadium (V), chromium (Cr), iron (Fe), and nickel (Ni), plus singly ionized Ti. Their Q coefficients have been calculated previously (*10*). The 17 pairs of lines have a wide range of sensitivity to α variation, with differences in Q within each pair from −0.08 to +0.18 (*10*).

We measured pair separations using a fully automated process for all of the 423 HARPS exposures. In each exposure, the core of each line—the central seven pixels, spanning ≈5.7 km s^{-1}—was fitted with a Gaussian model to determine the centroid wavelength. We then computed the wavelength differences between line pairs in each exposure, incorporating the corrections for the effects discussed above (*11*–*13*). We denote these pair separations $\Delta v^i_{\rm raw}$ for pair i. In principle, they can be compared to gauge any α variation between these 17 solar twins. However, an analysis of 130 stars spanning a larger range in $T_{\rm eff}$, [Fe/H], and log g (300 K, 0.3 dex, and 0.4 dex around solar values, respectively) has shown that pair velocity separation varies systematically with the stellar parameters, typically by ~60 m s^{-1} across this range (*12*, *13*). We fitted a quadratic model to those correlations and used it to compute the expected line pair separation for each star in our sample, denoting the resulting values $\Delta v^i_{\rm model}$. We also incorporated an intrinsic star-to-star scatter, $\sigma^i_{**} \approx 0$ to 15 m s^{-1} (*11*, *13*). We then used $\Delta v^i_{\rm model}$ to correct the observed separations for each individual star:

$$\Delta v^i_{\rm sep} = \Delta v^i_{\rm raw} - \Delta v^i_{\rm model}(T_{\rm eff}, [{\rm Fe/H}], \log g) \quad (2)$$

For each line pair i, the value of σ^i_{**} is the systematic error in $\Delta v^i_{\rm sep}$; it is the typical absolute value of the intrinsic deviation from the model.

The $\Delta v^i_{\rm sep}$ values have previously been calculated (*11*) from the HARPS solar twin exposures. For each line pair in each solar twin, the velocity separation measurements from multiple exposures were combined by using a weighted mean, with outliers excluded through an iterative process (*12*, *13*). Multiple exposures were available for 14 of the solar twins, allowing us to check for systematic errors as a function of time. The optical fibers that feed light from the telescope into HARPS were changed in mid-2015, resulting in large calibration changes. Analysis of the pre- and post–fiber change epochs separately—including the determination of $\Delta v^i_{\rm model}$—has shown no evidence for systematic differences in $\Delta v^i_{\rm sep}$ between them (*11*). We therefore combined their weighted mean $\Delta v^i_{\rm sep}$ values. Three of the 17 line pairs appear twice in each exposure because they are in the overlapping wavelength ranges of neighboring diffraction orders. We treat these two instances separately because we found differences of ~20 m s^{-1} between their $\Delta v^i_{\rm raw}$ values, which we ascribe to optical distortions within HARPS. Nevertheless, their weighted mean $\Delta v^i_{\rm sep}$ values show no systematic differences for our 17 stars or the larger sample (*12*), so we combined the $\Delta v^i_{\rm sep}$ value for two instances of a pair using a weighted mean.

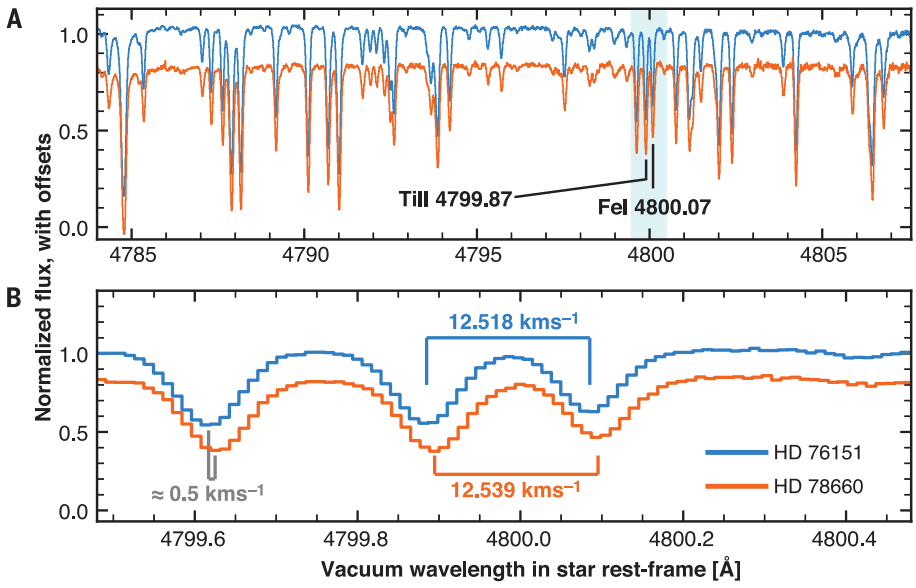


Fig. 1. Example solar twin spectra. (**A**) Small sections of continuum-normalized HARPS spectra of two solar twins from our sample, HD 76151 (blue) and HD 78660 (orange), with the latter shifted down by 0.2 for clarity. Black labels indicate an example line pair used to constrain α variation. This is the least separated pair of the 17 in our analysis; the maximum separation of 800 km s^{-1} would span approximately half the width of the figure. The cyan shading indicates the region shown in (B). (**B**) The measured velocity separations $\Delta v^i_{\rm raw}$, indicated with brackets, differ by 21 m s^{-1}, before correction for their stellar parameters (Eq. 2). In (A) and (B), the spectra are shown in the stars' rest frames; errors in their radial velocities are evident as a ≈0.5 km s^{-1} shift between them [(B), gray]. Our differential approach is insensitive to that offset.

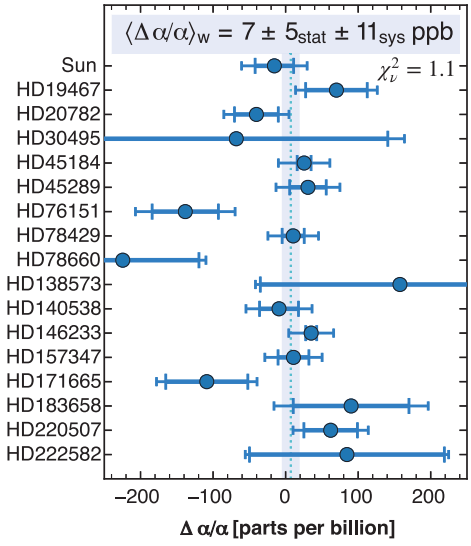


Fig. 2. Fine-structure constant measurements. The relative deviation of the fine-structure constant $\Delta\alpha/\alpha$ is shown for each star in our sample. These values are averages of all 17 line pairs for each star. The inner large error bars indicate the 1σ statistical uncertainties, dominated by the number of observations available of each star, whereas the outer small error bars combine the statistical and systematic uncertainties in quadrature. The weighted mean of the sample $\langle\Delta\alpha/\alpha\rangle_{\rm w}$ is indicated with the dotted cyan line; the blue shaded region indicates the combined 1σ statistical and systematic uncertainties. $\chi^2_\nu = 1.1$ is the reduced χ^2 (per degree of freedom) of the individual measurements around the weighted mean.

Our derived values of $\Delta\alpha/\alpha$ for each star are shown in Fig. 2. The $\Delta v^i_{\rm sep}$ value for a line pair i is converted to $\Delta\alpha/\alpha$ by using Eqs. 1 and 2 and the Q coefficient calculations (*10*). For each star in Fig. 2, the $\Delta\alpha/\alpha$ values from all pairs were consistent with each other, so they were combined by using a weighted mean. The weights in that process and the final uncertainties include the statistical uncertainties, derived from the SNR of the HARPS spectra, and systematic errors that incorporate the star-

to-star scatters for all line pairs (σ^i_{**}) and a smaller contribution from the uncertainties in the Q coefficients. Because a line can be shared by multiple pairs, its statistical and systematic uncertainties cause correlated errors across those pairs; we used a Monte Carlo method to compute the combined $\Delta\alpha/\alpha$ value and its statistical and systematic uncertainty for each star (*13*).

We found no variations in α between nearby solar twins (<50 parsec), with a typical (median) uncertainty in $\Delta\alpha/\alpha$ of ≈50 parts per billion (ppb) (adding statistical and systematic errors in quadrature). The precision reaches ≈30 ppb for some stars, which is ≳30 times more precise than individual quasar absorption systems (*2*, *4*). The systematic error term dominates in these cases (Fig. 2), mainly because of the intrinsic star-to-star scatter, $\sigma^i_{**} \approx 0$ to 15 m s^{-1} per line pair i. The solar twins method provides ≳100 times more accuracy than comparison between lines in individual white dwarfs or giant stars with their laboratory counterparts (*5*, *6*). The results for the 17 stars are formally consistent with each other, with $\chi^2 = 18.2$ around their weighted mean (16 degrees of freedom; 31% probability of a larger value by chance alone), so there is no evidence for additional systematic errors that are not accounted for by σ^i_{**}.

Another study (*12*) considered a variety of astrophysical and instrumental effects that could cause spurious variation of α between stars and/or account for σ^i_{**}. Apart from those already corrected in our analysis (such as wavelength calibration distortions), that study ruled out systematic error contributions from line blending; pair separation; differences in line depth in a pair; transiting exoplanets or magnetic activity cycles of the target stars; and contamination of spectra by scattered moonlight, cosmic ray events, or charge transfer inefficiencies in the detector. However, they estimated that variations in stellar rotation velocities or elemental and isotopic abundances between stars may plausibly explain the size of and variations in σ^i_{**} (*12*). Nevertheless, they did not find specific evidence for these effects with simple tests, even among the larger data sample of stars used (*12*).

Combining the results from all 17 stars provided a weighted mean with 12 ppb ensemble precision:

$$\langle \Delta\alpha/\alpha \rangle_w = 7 \pm 5_{stat} \pm 11_{sys} \text{ ppb} \qquad (3)$$

where $\langle \Delta\alpha/\alpha \rangle_w$ and its 1σ statistical uncertainty and systematic error were calculated from the Monte Carlo simulations to account for the correlations between results for different stars because they share common line pairs (*13*). The weights are the inverse variances from quadrature addition of the statistical and systematic uncertainties in $\Delta\alpha/\alpha$ for each star. The combined result (Eq. 3) acts as an entirely empirical reference for stellar measurements of α. This and the ability of our automatic analysis procedure to recover shifts in α between stars were tested by altering the wavelength measurements for half our twins by amounts corresponding to an α variation of 100 ppb (*13*). Rerunning the full analysis but removing these stars from the determination of Δv^i_{model}, we recovered an 86 ± 19 ppb difference between the shifted and unshifted twins. The discrepancy arises because some measurements of shifted lines are excluded as outliers; the shifts introduced are much larger than the total uncertainties (including σ^i_{**}). This confirms that our analysis process would still have detected any large (~100 ppb) discrepancies between some twins if they were present in the data.

REFERENCES AND NOTES

1. R. Lange *et al.*, *Phys. Rev. Lett.* **126**, 011102 (2021).
2. S. M. Kotuš, M. T. Murphy, R. F. Carswell, *Mon. Not. R. Astron. Soc.* **464**, 3679–3703 (2017).
3. M. T. Murphy, K. L. Cooksey, *Mon. Not. R. Astron. Soc.* **471**, 4930–4945 (2017).
4. M. T. Murphy *et al.*, *Astron. Astrophys.* **658**, A123 (2022).
5. A. Hees *et al.*, *Phys. Rev. Lett.* **124**, 081101 (2020).
6. J. Hu *et al.*, *Mon. Not. R. Astron. Soc.* **500**, 1466 (2021).
7. V. A. Dzuba, V. V. Flambaum, J. K. Webb, *Phys. Rev. Lett.* **82**, 888–891 (1999).
8. D. Dravins, *Annu. Rev. Astron. Astrophys.* **20**, 61–89 (1982).
9. J. I. González Hernández *et al.*, *Astron. Astrophys.* **643**, A146 (2020).
10. V. A. Dzuba, V. V. Flambaum, M. T. Murphy, D. A. Berke, *Phys. Rev. A* **105**, 062809 (2022).
11. D. A. Berke, M. T. Murphy, C. Flynn, F. Liu, *Mon. Not. R. Astron. Soc.* 10.1093/mnras/stac2458 (2022).
12. D. A. Berke, M. T. Murphy, C. Flynn, F. Liu, *Mon. Not. R. Astron. Soc.* 10.1093/mnras/stac2037 (2022).
13. Materials and methods are available as supplementary materials.
14. M. Mayor *et al.*, *Messenger* **114**, 20 (2003).
15. T. Wilken *et al.*, *Mon. Not. R. Astron. Soc.* **405**, L16–L20 (2010).
16. D. Milaković, L. Pasquini, J. K. Webb, G. Lo Curto, *Mon. Not. R. Astron. Soc.* **493**, 3997–4011 (2020).
17. A. Coffinet, C. Lovis, X. Dumusque, F. Pepe, *Astron. Astrophys.* **629**, A27 (2019).
18. J. W. Brault, *Mikrochim. Acta* **93**, 215–227 (1987).
19. M. Laverick *et al.*, *Astron. Astrophys.* **612**, A60 (2018).
20. D. A. Berke, DBerke/varconlib: VarConLib initial release. Zenodo (2022); doi:10.5281/zenodo.7196771.
21. D. A. Berke, M. T. Murphy, C. Flynn, F. Liu, DBerke/Berke_et_alia_2022_supplemental_data: Initial release of supplementary material. Zenodo (2022); doi:10.5281/zenodo.7196796.
22. M. T. Murphy, MTMurphy77/alpha_SolarTwins22: Code and data for constraining variations in the fine-structure constant between solar twin stars. Zenodo (2022); doi:10.5281/zenodo.7196515.

ACKNOWLEDGMENTS

We thank D. Dravins for discussions about potential astrophysical systematic errors. **Funding:** M.T.M., F.L., and C.L. acknowledge the support of the Australian Research Council through Future Fellowship grant FT180100194. V.A.D. and V.V.F. acknowledge the Australian Research Council for support through grants DP190100974 and DP200100150. OzGrav is funded by the Australian government through the Australian Research Council Centres of Excellence funding scheme. **Author contributions:** M.T.M. conceived the stellar twins technique, acquired funding, supervised the project, calculated the $\Delta\alpha/\alpha$ values, and wrote the draft manuscript. D.A.B. wrote the software, performed the velocity shift measurements and analysis, and curated all data. F.L., C.F., and C.L. assisted with the methodology and analysis and validation of results. V.A.D. and V.V.F. calculated the sensitivity coefficients (Q). All authors commented on and revised the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** This work is based on observations obtained from the ESO Science Archive Facility and collected at the European Southern Observatory under ESO program(s) listed in table S1. The observations are available from the ESO Science Archive facility: http://archive.eso.org/eso/eso_archive_main.html. Our software for measuring the line wavelengths and computing the line pair separations and models is available on Zenodo (*20*). Tables of the lines used in this study; their laboratory wavelengths; our measured and model offsets from those values; and our measured line pair separations, models, and σ^i_{**} values are available on Zenodo (*21*). Our software for computing $\Delta\alpha/\alpha$ in this work is available on Zenodo (*22*). The stellar parameters and our measured $\Delta\alpha/\alpha$ values and uncertainties are listed in table S2. **License information:** Copyright © 2022 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abi9232
Materials and Methods
Figs. S1 to S4
Tables S1 to S3
References (*23*–*34*)

Submitted 1 April 2022; accepted 14 October 2022
10.1126/science.abi9232

FLEXIBLE ELECTRONICS

Universal assembly of liquid metal particles in polymers enables elastic printed circuit board

Wonbeom Lee[1]†, Hyunjun Kim[1]†, Inho Kang[2], Hongjun Park[3], Jiyoung Jung[4], Haeseung Lee[1], Hyunchang Park[1], Ji Su Park[1], Jong Min Yuk[1], Seunghwa Ryu[4], Jae-Woong Jeong[2], Jiheong Kang[1]*

An elastic printed circuit board (E-PCB) is a conductive framework used for the facile assembly of system-level stretchable electronics. E-PCBs require elastic conductors that have high conductivity, high stretchability, tough adhesion to various components, and imperceptible resistance changes even under large strain. We present a liquid metal particle network (LMP_{Net}) assembled by applying an acoustic field to a solid-state insulating liquid metal particle composite as the elastic conductor. The LMP_{Net} conductor satisfies all the aforementioned requirements and enables the fabrication of a multilayered high-density E-PCB, in which numerous electronic components are intimately integrated to create highly stretchable skin electronics. Furthermore, we could generate the LMP_{Net} in various polymer matrices, including hydrogels, self-healing elastomers, and photoresists, thus showing their potential for use in soft electronics.

Stretchable electronics with high stretchability and high toughness are essential for soft robotics (*1*, *2*), skin electronics (*3*, *4*), and implantable electronics (*5*, *6*). Substantial progress has been made in intrinsically stretchable conductors. High metallic conductivity with rubber-like stretchability has been successfully achieved in conductive polymers (*7*) and nanocomposites (*8*–*12*). However, certain critical challenges, including the inevitable change in electrical resistance during stretching and difficulty in achieving long-term cyclic stability and strong interfacial bonding with electronic components, remain. Hence, an elastic printed circuit board (E-PCB) has not been realized without structure engineering (*13*).

Room-temperature liquid metals (LMs) have received considerable attention as elastic conductors because of their metallic conductivity and extreme deformability. Gallium-based LMs have been studied for elastic conductors through various approaches, including dispersing LM particles (LMPs) in an elastomer (*14*, *15*), coating LM on a porous polymer matrix (*16*), mixing LMPs with a solid conductive filler composite (*17*, *18*), doping LMPs in a polymer matrix (*19*), and forming a biphasic LM structure (*20*). However, LM-based conductors suffer from leakage issues under external mechanical stimuli, which limit their reliability, uniformity, and stability.

[1]Department of Materials Science and Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea. [2]School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea. [3]Center for Nanomaterials and Chemical Reactions, Institute for Basic Science (IBS), Daejeon 34141, Republic of Korea. [4]Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea.
*Corresponding author. Email: jiheongkang@kaist.ac.kr
†These authors contributed equally to this work.

We report a universal synthetic route for highly conductive and mechanically tough LMP-based conductors without the LM leakage issue (Fig. 1). Our LM conductor includes a long-range assembled network of LMPs (LMP_{Net}) and a tough elastomeric matrix. The percolation structure and deformation mechanisms of the LMP_{Net} led to high conductivity, outstanding toughness, and imperceptible resistance changes under large deformation (Fig. 1C). The LMP_{Net} is composed of a network of large LMPs (average size of 2 to 3 μm) as the main framework and smaller LMPs (average size of 100 nm, denoted as LMP_{nano}) as network interconnectors (Fig. 1B and figs. S1 to S3). When the LMP_{Net} is stretched, the micrometer-sized LMPs deform into ellipsoidal structures, whereas the LMP_{nano} interconnectors remain intact similarly to solid particles (Fig. 1C and figs. S1 and S2). Thus, particle–particle contacts could be preserved under large strain, resulting in negligible resistance changes under large strain (>4000%).

The LMP_{Net} could be formed with high uniformity and reliability over a large area (fig. S4). This allowed us to predict the electrical resistance of the lines and design an E-PCB, in which various electronic components, including integrated circuit (IC) chips, resistors, transistors, and capacitors are assembled with tough interfacial adhesion (Fig. 1D).

As illustrated in Fig. 1, A and B, and fig. S5, the formation of the LMP_{Net} in a polymer matrix was accomplished in two steps: First, an LMP–polymer composite was formed, and then the LMP_{Net} formation was induced. To prepare micrometer-sized LMPs, we applied an acoustic field to the room-temperature LM alloy (eutectic gallium indium, EGaIn) (1.35 g) in acetone (30 ml) at an amplitude of 63.4 μm (44%) for 20 min, using a probe sonicator in water. The solvent was decanted after centrifugation for 30 min at 2200 rpm. The sedimented LMPs were mixed with a thermoplastic elastomer, polyurethane (PU) dissolved in dimethylacetamide (DMAc, 200 mg/ml), by using a THINKY mixer for 10 min. Scanning electron microscopy images and the size distribution of the LMP samples processed by various acoustic field application times are shown in fig. S6. Thereafter, the LMP–PU ink was printed on a PU substrate, and the printed lines were annealed at 80°C for 24 hours to completely remove the residual solvent (fig. S7).

The as-printed lines are insulating because of the large interparticle distance of LMPs in the polymer matrix that is induced by the electrostatic repulsion of native oxides on LMPs (*21*) (Fig. 1B, left). Generally, LMP-based conductors require an activation (sintering) step to interconnect the LMPs and achieve high conductivity. Various activation methods, such as external mechanical force application (*14*, *15*), high-temperature sintering (*20*), and laser sintering (*22*), have been reported. These methods rupture the native oxide of LMPs and cause a substantial amount of LM to leak to other LMPs, thereby forming percolation pathways. These methods have limitations in the reliability, uniformity, and mechanical robustness of circuit lines. During the activation, a large amount of LM emerges from the surface of conductive lines and disrupts other lines (fig. S8).

We devised an alternative method for forming a highly conductive LMP assembled network in the polymer matrix without LM leakage. This method involves acoustic field application to as-printed LMP lines. For this, we chose water as the medium to avoid unwanted damage to the printed lines or temperature increase. When an acoustic field was applied for 30 s, we observed the formation of nanosized LMP (LMP_{nano}) between the original micrometer-sized LMPs (Fig. 1B and figs. S1 to S3), resulting in the interconnection of the LMPs with LMP_{nano}, that is, the formation of the LMP_{Net}.

We investigated the electrical characteristics of the LMP_{Net} depending on the LM volume fraction in the composite. The as-printed LMP conductor was insulating, even at a high LM content, and required an acoustic field to form a conductive LMP_{Net} (Fig. 2A). The conductivity of the LMP_{Net} increased with an increase in the LM volume fraction in the composite, and it reached 2.10×10^6 S/m at 74.4 vol % of the LM (Fig. 2A and fig. S9). Considering the volume fraction of the polymer in the composite and the electrical conductivity of the pure LM, the resultant conductivity nearly reached the theoretical effective conductivity calculated by using the relationship $\sigma = \sum f_A \sigma_A$, where f is the volume fraction of the conductive material and σ is the electrical conductivity of the conductive component (Fig. 2A and fig. S10) (*23*). This result suggests that a nearly

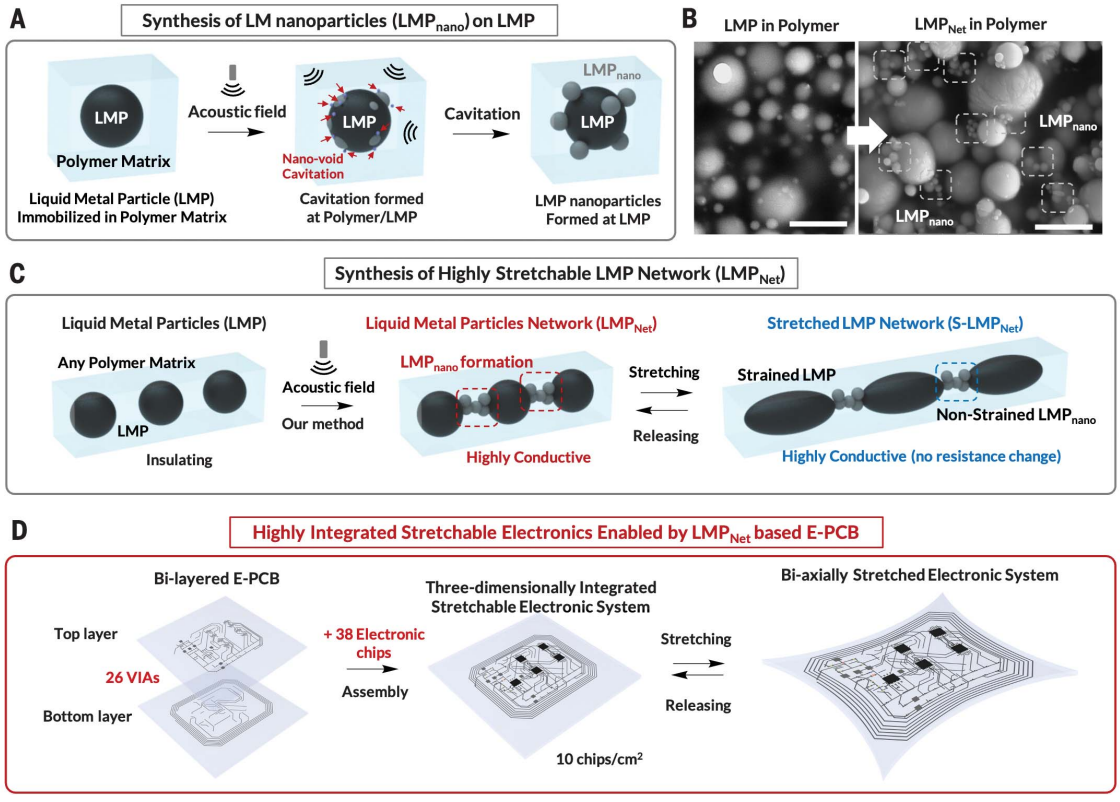


Fig. 1. Formation of a liquid metal particle network in a polymer and its application for elastic printed circuit boards. (**A**) Schematic of the formation of liquid metal nanoparticles (LMP$_{nano}$) at the surface of existing micrometer-sized liquid metal particles (LMPs) via acoustic field application. (**B**) Scanning electron microscopy images of the LMPs and LMP network (LMP$_{Net}$) formed in the polymer, which reveal that LMP$_{nano}$ species are formed at the surface of the original LMPs. Scale bars, 3 μm. (**C**) Schematic illustration of the formation of a highly conductive LMP$_{Net}$ and the distinctive behavior of the elastic LMP$_{Net}$ derived from the differences in the sizes of the original LMP and LMP$_{nano}$, resulting in negligible resistance change during stretching. (**D**) Schematic illustration of the elastic multilayered printed circuit board based on the LMP$_{Net}$ and the assembly of integrated stretchable electronics.

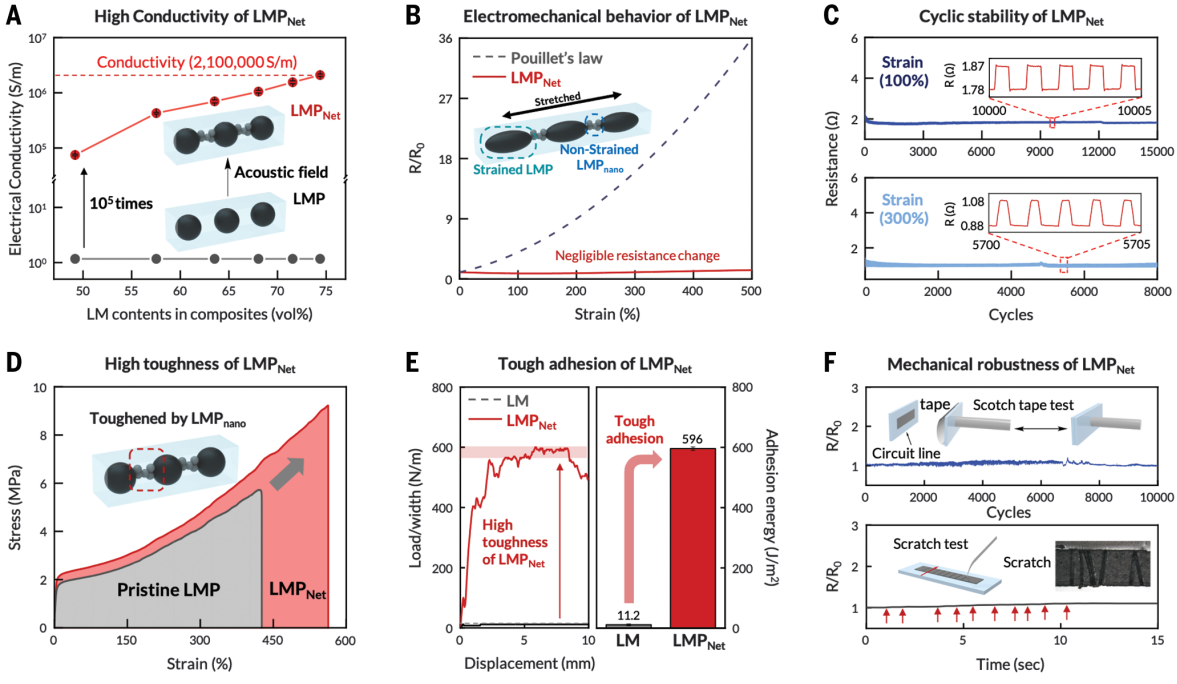


Fig. 2. Electrical and mechanical properties of LMP$_{Net}$. (**A**) Electrical conductivity of the LMP–polymer composite as a function of the liquid metal content before and after the LMP$_{Net}$ formation; the electrical conductivity of the LMP$_{Net}$ reached 2.10 × 10^6 ± 9.8 × 10^4 S/m (n = 3, where n is the number of samples that are used to generate statistical data). (**B**) Relative resistance changes of the LMP$_{Net}$ under uniaxial strain of up to 500% (red line, R/R_0 = 1.33), and theoretical prediction based on an incompressible spherical conductor [gray dot, $R/R_0 = (1+\epsilon)^2$, where ϵ is the applied strain]. (**C**) Relative resistance changes of the LMP$_{Net}$ line during 15,000 cycles of stretching at 100% strain (top, dark blue) and 8000 cycles of stretching at 300% strain (bottom, pale blue). The insets show a detailed resistance response of the conductor to the applied strain. (**D**) Stress–strain curves of the LMP–polymer (gray) and LMP$_{Net}$–polymer (red) composites. The mechanical strength increased by 160% and the toughness increased by 190% when the LMP$_{Net}$ was formed. (**E**) Interfacial adhesion strength of the pure LM and LMP$_{Net}$ on a surface-functionalized substrate. The interfacial adhesion strength of the LMP$_{Net}$ (596 ± 5.6 J/m^2, n = 3, where n is the number of samples that are used to generate statistical data) on the surface-functionalized substrate is more than 5300% higher than that of the pure LM (11.2 ± 3.2 J/m^2, n = 3, where n is the number of samples that are used to generate statistical data). (**F**) Relative resistance changes of the LMP$_{Net}$ under external stimuli, Scotch-taping (top), and scratching (bottom), which demonstrate the mechanical and electrical robustness of the LMP$_{Net}$.

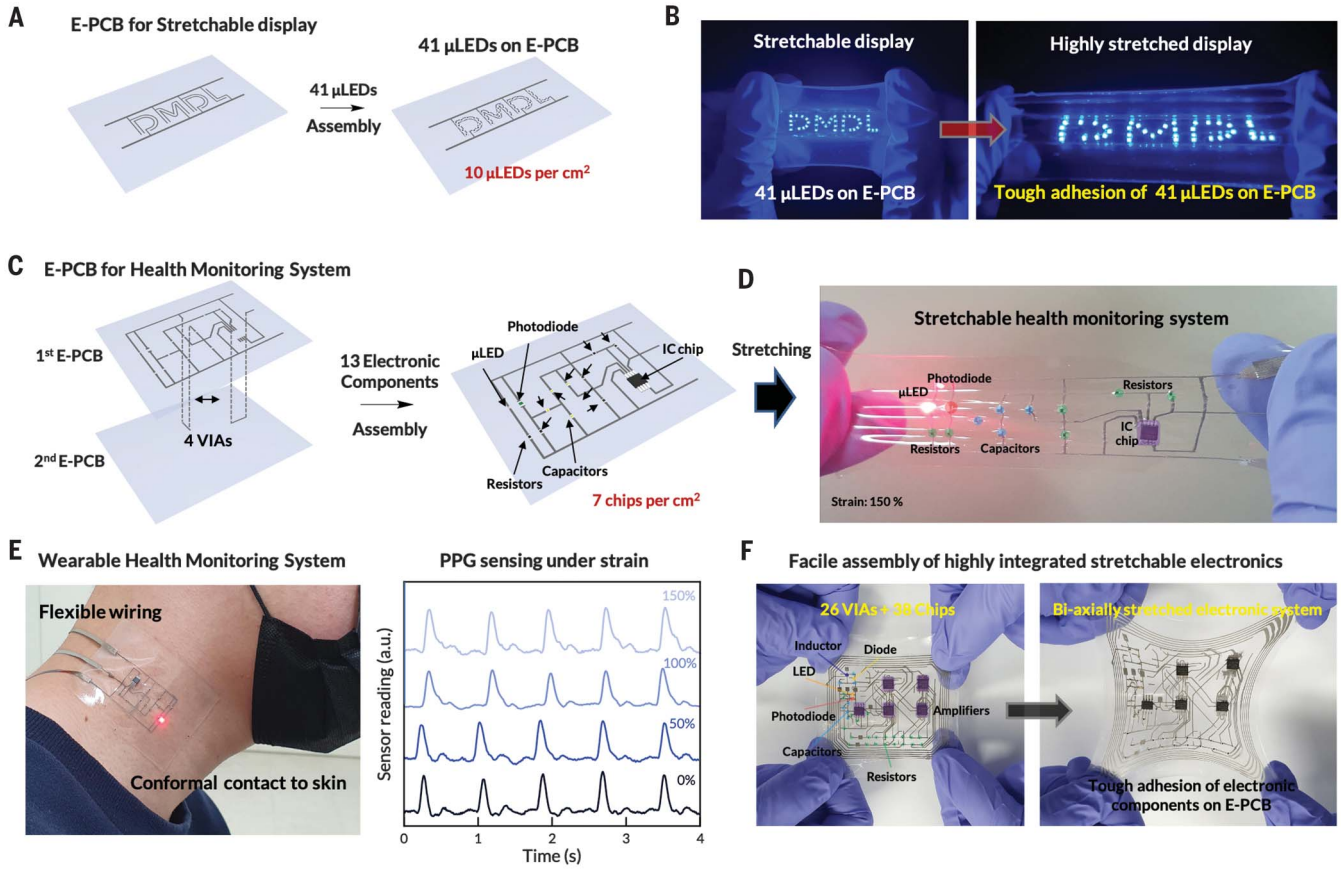


Fig. 3. LMP_{Net}-based E-PCB. (**A**) Schematic illustration of the circuit line and an LED array assembled using the LMP_{Net} for a stretchable display. (**B**) Optical images of the LED array reading "DMDL" at rest (left) and under 100% strain (right). (**C**) Schematic illustration of the circuit lines, VIAs, and electronic components assembled using the LMP_{Net} for a stretchable health monitoring system. (**D**) Optical image showing the operation of a photoplethysmography (PPG) sensor before and after stretching by 150%. (**E**) Optical image of the PPG sensor attached to human skin (left) and the normalized reading of the PPG sensor as a function of the strain (0, 50, 100, and 150% strain) applied to the PPG sensor circuit board (right). (**F**) Optical image of a stand-alone integrated stretchable electronic system (left) and biaxially stretched electronic system (right). This system includes 26 VIAs and 38 electronic components.

defect-free percolation pathway was formed by the assembly of the LMPs in the polymer matrix. Although the LMP_{Net} is based on a zero-dimensional particle structure, its electrical conductivity is higher than those of the previously reported elastic printable conductors (fig. S11) (*7*, *9*, *10*, *16*–*18*, *20*). The LMP_{Net} exhibited zero resistance change at 100% strain (R/R_o = 1.00) and excellent electromechanical decoupling at 600, 2000, and 4100% strain (R/R_o = 1.41, 5.18, and 20.8, respectively) (Fig. 2B and fig. S12). It has excellent environmental stability (phosphate-buffered saline, 16 weeks), thermal stability (fig. S13), and high cyclic stability (15,000 cycles at 100% strain, 8000 cycles at 300% strain, and 1200 cycles at 500% strain) (Fig. 2C and fig. S14). We investigated the effects of the strain rate and LM volume fraction on the electromechanical behavior of the LMP_{Net} conductor. The LMP_{Net} exhibits a stable electrical performance during dynamic stretching, regardless of the strain rate (fig. S15) and the LM volume fraction (fig. S16). We also confirmed its imperceptible change in the resistance even under biaxial stretching (fig. S17).

According to Eshelby's theory, LMPs in an elastomeric matrix deform their structures when a strain is applied. This feature allows much better electromechanical responses from LMP-based conductors than from conventional rigid conductive filler-based conductors (fig. S18). The electromechanical properties of our LMP_{Net} conductor result from not only the deformation of the LMPs but also the structure of the assembled LMP network. The deformation of LMPs is size dependent, with nanosized LMP_{nano} mimicking the behavior of solid particles (*24*). We performed theoretical simulations and confirmed this size-dependent deformation of LMPs in the polymer (fig. S19) (*24*–*26*). When LMP_{Net} was stretched, we observed that large LMPs deformed to ellipsoidal structures and the LMP_{nano}'s between large LMPs were intact (fig. S1). The dissimilar deformation of the LMPs of two different sizes enables the realization of a mechanically and electrically resilient percolation network (Fig. 1C).

To understand the impact of the LMP_{Net} on the mechanical properties, we performed mechanical tests on three samples: pure PU, LMP–PU, and LMP_{Net}–PU. In contrast with rigid conductive fillers, LMP inclusions in PU, which have zero Young's modulus, make the LMP–PU composite softer (fig. S20). However, the deformation and crack-bridging effect of the LMP substantially toughen the polymer matrix (*27*). We investigated how the LMP_{nano} affects the mechanical properties of the composite. As discussed above, LMP_{nano} behaves like a solid particle in PU and thus has a stiffening effect on the PU matrix (Fig. 2D and fig. S21). LMP_{nano} also improved the stretchability and toughness of the composite (Fig. 2D) because of the additional energy dissipation mechanism facilitated by the reversible assembly of the LMP_{Net} (fig. S22).

The interfacial adhesion problem of LMP conductors with rigid electronic chips is severe in E-PCB. Owing to the low toughness of LMP, electronic chips are easily delaminated from conductive lines by cohesive failure when stretched.

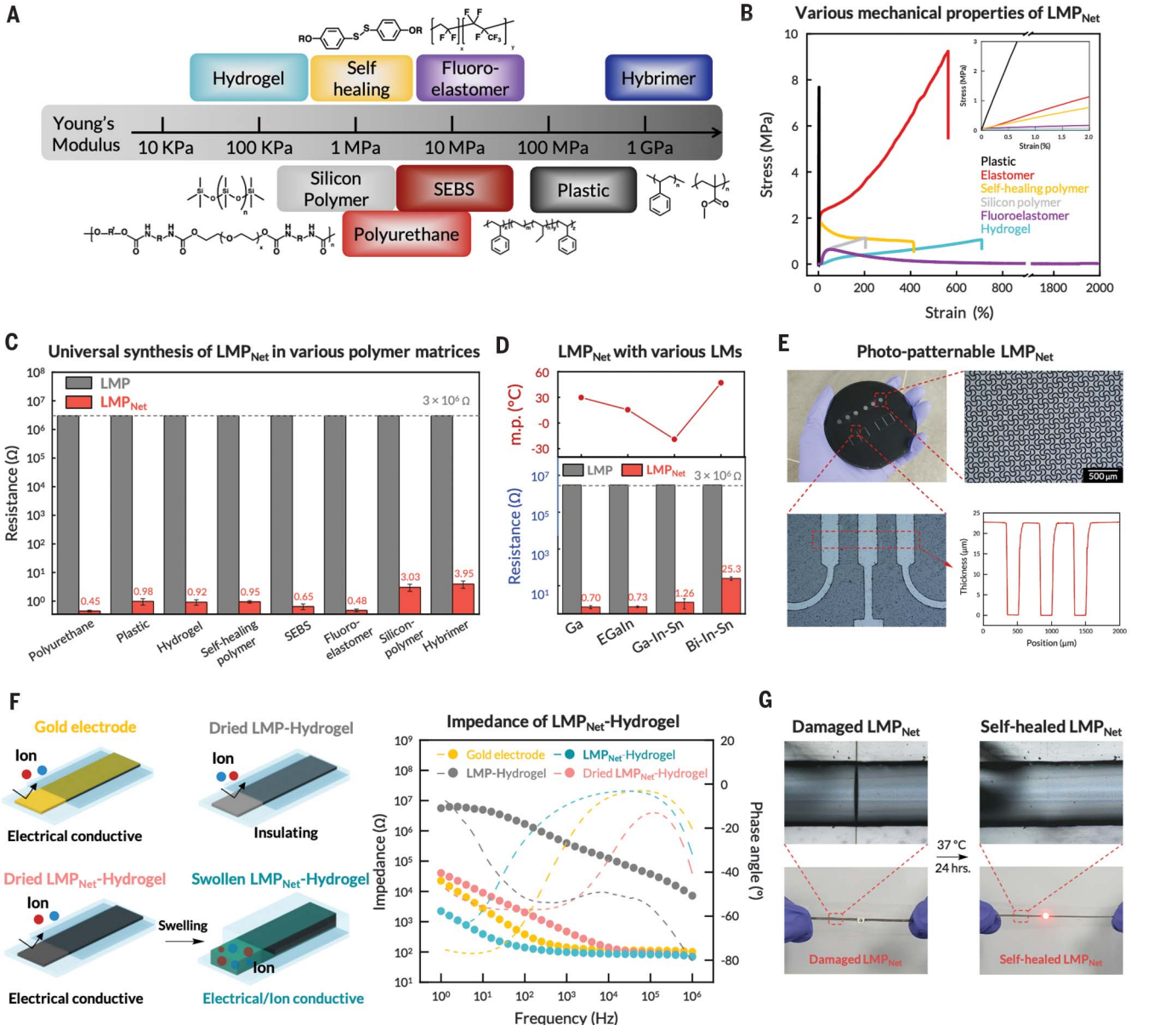


Fig. 4. LMP_{Net} in various polymers. (**A**) Young's modulus of various polymer matrices used for LMP_{Net} formation. (**B**) Stress–strain curves of the LMP_{Net} formed in various polymers (plastic, elastomer, self-healing polymer, silicon-based polymer, fluoropolymer, and hydrogel); the data reveal that the mechanical properties can be tuned by choosing an appropriate polymer. (**C**) Resistance of the LMP–polymer and LMP_{Net}–polymer composites based on diverse polymer matrices. The LMP_{Net} shows low electrical resistance (<5 ohms), regardless of the host matrix (n = 4, where n is the number of samples that are used to generate statistical data). (**D**) Melting point (top) and resistance (bottom) of the LMP_{Net} formed from four types of liquid metal alloys with different compositions; the data demonstrate that various types of metals with low melting points can also be used to form the LMP_{Net} (n = 4, where n is the number of samples that are used to generate statistical data). (**E**) Digital image (top left), optical microscopy image (top right, bottom left), and thickness profile (bottom right) of a photopatterned LMP_{Net}. (**F**) Schematic illustration of various electrodes (left) and their impedance spectra (right); gold electrode (yellow), dried LMP–hydrogel (gray), dried LMP_{Net}–hydrogel (red), and swollen LMP_{Net}–hydrogel (blue). Swollen LMP_{Net}–hydrogel exhibits the lowest impedance owing to the penetration of ions into the hydrogel (left). (**G**) Optical microscopy images (top) and digital images (bottom) of a damaged and self-healed LMP_{Net}.

Our LMP_{Net}–PU lines afford high interfacial adhesion energy (596 J/m^2) with various engineered surfaces. This can be attributed to the adhesion properties of PU and efficient energy dissipation facilitated by the LMP_{Net} (Fig. 2E and fig. S23). As demonstrated in fig. S24 and movie S1, a commercial micro–light-emitting diode (μLED) bonded to LMP_{Net}–PU lines exhibited stable performance under dynamic deformation even without the aid of an encapsulation layer.

One of the major challenges of the reported LMP-based conductors is the leakage of the LM owing to the continuous rupture of LMPs by mechanical stimuli. For example, upon scratching or stretching, several LM droplets emerged from the conductive lines (fig. S8). By contrast, no LM leakage could be observed from our LMP_{Net} conductive lines, and no degradation of the electrical properties was noted upon scotch-taping, scratching, or stretching the conductive lines (Fig. 2F and fig. S25). The mechanical robustness of the LMP_{Net} can be attributed to both the high toughness of the LMP_{Net}–PU and the relatively small size of the LMPs. In this study, we used LMPs with an average size of 2 μm as the main component of

the LMP_{Net}. In previously reported LMP systems (*14*, *15*, *19*), large LMPs (more than 10 μm) were used to achieve high conductivity. We experimentally confirmed the LM leakage issue in the case of an LMP_{Net} with LMP sizes of >5 μm (fig. S8). Notably, we achieved a high conductivity of the LMP_{Net} even with 2-μm LMPs (Fig. 2A).

To understand how LMP_{nano} is generated by an acoustic field, we studied multiple possible mechanisms (figs. S26 to S27). The acoustic field can generally be thought to cause a temperature increase or the collapse of nanobubbles on the surface of the printed lines (*28*). We did not observe any evolution of the percolation network when the LMP lines were annealed at 150°C for 1 day (fig. S28A), indicating that the LMP_{nano} formation is not due to temperature increase. Further, we could also exclude the effect of nanobubbles formed in the medium and their collapse at the water/line interface through two experiments. If the network formation is a surface event, a thickness dependence of the LMP_{Net} formation is expected. However, the LMP_{Net} could be generated even in a 60-μm-thick film with the same electrical conductivity as that of a 20-μm-thick film (fig. S29). We covered the LMP lines with a thick PU film and applied an acoustic field. We observed the successful formation of the LMP_{Net}, suggesting that it was not a surface event (fig. S28B).

We hypothesize that the LMP_{Net} formation occurs in three distinct steps (Fig. 1A and fig. S30). First, the acoustic energy is transferred from the probe sonicator to the composite via the water medium. Then, the acoustic energy accumulates at the LMP/polymer/LMP interface. As the polymer is acoustically transparent, the acoustic wave energy can freely travel and reach the bottom part of the polymer film (table S1). By contrast, LMPs have a high acoustic impedance and mostly reflect the wave energy. Owing to the large difference in acoustic impedance between the polymer and LMP, the wave energy mostly accumulates at the LMP/polymer/LMP interfaces. We confirmed this possibility through finite element simulation using a simple model (fig. S31). The accumulated acoustic energy starts to form nanobubbles (cavitation) at the LMP/polymer/LMP interface. The final step is the generation of LMP_{nano} from the original LMPs by the collapse of the nanobubbles (Fig. 1A). That is, LMP_{nano} is generated directly from the LMPs without complete rupture of LMPs. This event occurs inside the LMP–polymer composite, wherein the LMPs are immobilized. Therefore, LMP_{nano} forms efficiently and interconnects the LMPs without moving to other free spaces in the polymer matrix (fig. S1). Thus, a highly conductive LMP_{Net} was formed through these steps, in which large LMPs were compactly assembled by LMP_{nano} with an interparticle distance of less than 2 nm (fig. S3B). A similar outcome can be expected if LMP_{nano} is mixed with LMPs in the conductive ink. However, we could not observe any evolution of the percolation network, which indicated that LMP_{nano} did not interconnect the LMPs (insulating) (fig. S32).

To realize E-PCB, we confirmed the processability of the LMP_{Net} conductor. Our LMP_{Net} conductor provides an excellent platform via a printing process for the fabrication of E-PCB. We confirmed excellent printability, stretchable vertical interconnect accesses (VIAs), and chip-bonding processes (figs. S33 to S35). As a proof of concept, we fabricated bi-layer E-PCB and assembled integrated electronics, including a μLED array (Fig. 3, A and B; fig. S36; and movie S2) and a photoplethysmography (PPG) sensing wearable device (Fig. 3C and fig. S37). As shown in Fig. 3, B, D, and E, they can exhibit stable electrical performance under dynamic stretching. In addition, we fabricated highly integrated skin electronics, in which 26 VIAs and 38 chips were assembled, and confirmed their robustness under biaxial stretching (Fig. 3F and fig. S38).

Our acoustic field–based LMP_{Net} synthesis should apply to most polymer matrices. We successfully formed LMP_{Net} in more than 15 different polymers with various chemical and mechanical properties (Fig. 4, B and C; figs. S39 and S40; and table S2). We also confirmed the formation of LMP_{Net} using other LM alloys with different melting temperatures and compositions (Fig. 4D). All the LMP_{Net} systems exhibited high conductivity, regardless of the polymer and LM.

The LMP_{Net} could also be formed in a photoresist (SU-8), which enabled the high-resolution patterning of LMP_{Net} conductive lines (Fig. 4E and fig. S41). We observed unusual electrochemical features in a LMP_{Net}–hydrogel. We obtained the impedance spectrum of the LMP_{Net}–hydrogel in the frequency range 1 to 1×10^6 Hz and observed that our LMP_{Net}–hydrogel electrode exhibits one order of magnitude lower impedance than Au electrodes in the low-frequency range (Fig. 4F and figs. S42 to S44). This observation results from the high dual conductivity (electronic and ionic conduction) and high surface area of the LMP_{Net} in the hydrogel. The initial dry LMP_{Net}–hydrogel was only electronically conductive, whereas the sample swollen with electrolyte solution exhibited mixed electron and ion conduction, leading to extremely low impedance (*6*). Considering the biocompatibility of EGaIn and hydrogel matrix, our LMP_{Net}–hydrogel can be designed as an implantable electrode. Moreover, we formed an LMP_{Net} in a self-healing elastomer (SHE) (Fig. 4C) and investigated its mechanical and electrical self-healing properties (*29*, *30*). The LMP_{Net}–SHE could autonomously restore its mechanical and electrical properties when damaged (Fig. 4G).

REFERENCES AND NOTES

1. S. I. Rich, R. J. Wood, C. Majidi, *Nat. Electron.* **1**, 102–112 (2018).
2. D. Rus, M. T. Tolley, *Nature* **521**, 467–475 (2015).
3. D.-H. Kim *et al.*, *Science* **333**, 838–843 (2011).
4. S. Wang *et al.*, *Nature* **555**, 83–88 (2018).
5. D.-H. Kim *et al.*, *Nat. Mater.* **9**, 511–517 (2010).
6. Y. Liu *et al.*, *Nat. Biomed. Eng.* **3**, 58–68 (2019).
7. Y. Wang *et al.*, *Sci. Adv.* **3**, e1602076 (2017).
8. M. Park *et al.*, *Nat. Nanotechnol.* **7**, 803–809 (2012).
9. N. Matsuhisa *et al.*, *Nat. Mater.* **16**, 834–840 (2017).
10. T. Sekitani *et al.*, *Nat. Mater.* **8**, 494–499 (2009).
11. S. Choi *et al.*, *Nat. Nanotechnol.* **13**, 1048–1056 (2018).
12. Y. Kim *et al.*, *Nature* **500**, 59–63 (2013).
13. N. Matsuhisa, X. Chen, Z. Bao, T. Someya, *Chem. Soc. Rev.* **48**, 2946–2966 (2019).
14. E. J. Markvicka, M. D. Bartlett, X. Huang, C. Majidi, *Nat. Mater.* **17**, 618–624 (2018).
15. A. Fassler, C. Majidi, *Adv. Mater.* **27**, 1928–1932 (2015).
16. Z. Ma *et al.*, *Nat. Mater.* **20**, 859–868 (2021).
17. K. Parida *et al.*, *Nat. Commun.* **10**, 2158 (2019).
18. J. Wang *et al.*, *Adv. Mater.* **30**, e1706157 (2018).
19. S. Veerapandian *et al.*, *Nat. Mater.* **20**, 533–540 (2021).
20. S. Liu, D. S. Shah, R. Kramer-Bottiglio, *Nat. Mater.* **20**, 851–858 (2021).
21. Y. Lin *et al.*, *Nanoscale* **10**, 19871–19878 (2018).
22. S. Liu *et al.*, *ACS Appl. Mater. Interfaces* **10**, 28232–28241 (2018).
23. N. F. Uvarov, *Solid State Ion.* **136–137**, 1267–1272 (2000).
24. R. W. Style *et al.*, *Nat. Phys.* **11**, 82–87 (2015).
25. X.-Q. Zheng, H. Zhao, Z. Jia, X. Tao, P. X.-L. Feng, *Appl. Phys. Lett.* **119**, 013505 (2021).
26. G. M. Odegard, T. C. Clancy, T. S. Gates, *Polymer* **46**, 553–562 (2005).
27. N. Kazem, M. D. Bartlett, C. Majidi, *Adv. Mater.* **30**, e1706594 (2018).
28. K. S. Suslick, *Science* **247**, 1439–1445 (1990).
29. J. Kang, J. B.-H. Tok, Z. Bao, *Nat. Electron.* **2**, 144–150 (2019).
30. J. Kang *et al.*, *Adv. Mater.* **30**, e1706846 (2018).

ACKNOWLEDGMENTS

Funding: This study was supported by the National R&D Program funded by the Ministry of Science and ICT (grant no. NRF-2021M3H4A1A03048658), the Basic Science Research Program (grant no. NRF-2021R1C1C1011116), and the Wearable Platform Materials Technology Center (WMC, grant no. NRF-2022R1A5A6000846). This study was also partially supported by the National R&D Program (NRF-2021R1A4A1052070, NRF-2021M3H4A3A01050378, and NRF-2022M3E5E9017759) and 2020 Joint Research Project of Institutes of Science and Technology.
Author contributions: W.L., H.K., and J.K. conceived and designed the experiments. W.L. and H.K. prepared and characterized the LMP-based stretchable conductors. I.K. and J.-W.J. programmed the stretchable PPG sensor for signal conditioning circuit board demonstration. I.K. and W.L. fabricated the stretchable PPG sensor. W.L., H.K., H.P., H.L., J.S.P., and J.M.Y. performed scanning electron microscopy and transmission electron microscopy. J.J. and S.R. performed the computer simulations. W.L., H.K., and J.K. wrote the manuscript. All the authors discussed the results and commented on the manuscript. J.K. supervised this study.
Competing Interests: W.L., H.K., and J.K. are inventors in a patent application (Korean Patent application no. 10-2022-0036393, patent pending) that covers the formation of the LMP_{Net} in various polymers and their applications. **Data and materials availability:** All data are available in the main text or supplementary materials.

License information: Copyright © 2022 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.sciencemag.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abo6631
Materials and Methods
Figs. S1 to S44
Tables S1 and S2
References (*31*–*48*)
Movies S1 and S2

Submitted 17 February 2022; accepted 23 September 2022
10.1126/science.abo6631

PLANT MORPHOLOGY

Hydraulic failure as a primary driver of xylem network evolution in early vascular plants

Martin Bouda[1]*, Brett A. Huggett[2], Kyra A. Prats[3,4], Jay W. Wason[5], Jonathan P. Wilson[6], Craig R. Brodersen[3]*

The earliest vascular plants had stems with a central cylindrical strand of water-conducting xylem, which rapidly diversified into more complex shapes. This diversification is understood to coincide with increases in plant body size and branching; however, no selection pressure favoring xylem strand-shape complexity is known. We show that incremental changes in xylem network organization that diverge from the cylindrical ancestral form lead to progressively greater drought resistance by reducing the risk of hydraulic failure. As xylem strand complexity increases, independent pathways for embolism spread become fewer and increasingly concentrated in more centrally located conduits, thus limiting the systemic spread of embolism during drought. Selection by drought may thus explain observed trajectories of xylem strand evolution in the fossil record and the diversity of extant forms.

Water availability is a critical limiting factor for land plants (*1*), whose macroevolution is marked by a series of hydraulic milestones that mitigated water loss, provided control over transpiration, and increased water transport efficiency (*2*, *3*). These adaptations mark distinctions between major land plant lineages (*2*), each releasing plants from hydraulic constraints and thereby enabling them to expand their niche space into drier environments (*4*). The earliest tracheophytes had a simple cylindrical (terete), centrally located vascular strand [stele (*5*)] containing xylem with tracheids as the water conducting cells (*6*). This terete haplostele shape occurs repeatedly early in the fossil record, but steles soon diversified toward more radially elongated shapes or larger, more elaborate forms (Fig. 1A) (*3*, *5*, *6*). Despite a century-long debate over the evolutionary drivers of stele complexity from the Devonian and Carboniferous periods onward (*3*, *7*–*10*), no underlying selective pressure has been found to account for the observed patterns. Presently, the dominant view is that changes to vascular architecture were a developmental artifact of increasingly branched or complex plant bodies (*3*, *11*, *12*).

Coordination of the vasculature with branching or appendages is necessary for efficient plant body construction, and the ontogenetic connection between the two is well established (*12*). Nevertheless, developmental constraints imposed by the requirements of coordination across vascular traces present an insufficient explanation for many of the observed stelar forms. The extent of xylem strand medullation is commonly underdetermined by branching or attachment of vascular appendages, and, in many cases, there is no direct correspondence (*3*). The well-known correlation of increased medullation of the stele with plant size or axis diameter (*3*, *12*, *13*) led Bower (*7*) and Wardlaw (*8*) to hypothesize that the latter led to the former. Nevertheless, a physiologically accountable morphometric relationship has remained elusive (*3*, *10*, *14*).

Here, we propose the idea that stelar evolution may have increased drought resistance by altering xylem network topology. To replace evaporative losses from photosynthesizing surfaces, plants extract soil water under increasing capillary tension as soils dry. Sufficient tension places xylem water in a metastable state, making it increasingly vulnerable to cavitation, which blocks water transport locally with a vapor-phase embolus (*15*). Unchecked, the systemic spread of emboli leads to hydraulic network failure, tissue death (*16*), and ultimately plant mortality (*17*). Embolism spreads when the liquid tension in a water-filled conduit overcomes the air-seeding pressure (ASP) threshold of the pit membrane that separates it from an embolized one (*15*). Thus, conduit adjacency and connectivity make up a tissue-scale network, the topology of which determines the extent of embolism spread for a given ASP distribution (*18*, *19*). Network topology alone can push the vulnerability curve, measured as the cumulative proportion of xylem conductivity lost versus water tension, toward greater resistance to hydraulic failure (Fig. 1, B and C, and fig. S1) and thus shift the value of tension that constitutes a plant's drought mortality threshold [p88, defined as the xylem pressure leading to 88% loss of conductivity (*20*)].

Because tracheid network topology is determined by xylem strand shape and conduit packing, we propose that stele shape may have been subject to a drought-induced selective pressure. We first evaluate this hypothesis by quantifying the effect of network topology on simulated drought mortality thresholds in different xylem strand shapes. Starting from the terete haplostele characteristic of the earliest tracheophytes (*3*), we performed simulations with incrementally changing xylem strand shapes by introducing and gradually extruding lobing or a central pith (Fig. 2A, fig. S2, and movies S1 to S4). The resulting set of xylem strand shapes captures morphologies that are observable in the fossil record or in extant lycophytes and ferns [fig. S3; (*3*, *5*, *6*)]. Our model uses ASP distributions to determine the probability of embolism spread between adjacent conduits for a given level of drought stress. Because the magnitude of the topological effect scales with the variance in ASP, we report ASP data on seven representative extant species as a gauge of its biological importance (fig. S4A). By evaluating the sensitivity of the mortality threshold to stepwise topological changes, we effectively reconstruct the proposed dimension of the fitness landscape faced by land plants.

As xylem strand shape diverges from the terete haplostele, the mean number of neighbors per conduit decreases (Fig. 2). This progressively reduces the number of independent paths for embolism spread through the network. The remaining paths become concentrated in fewer, more topologically central conduits. We quantify path concentration as the squared root of the sum of squares of the number of shortest paths passing through each conduit divided by the total number of such paths (fig. S5). As the logarithm of path concentration (lnPC) increases, fewer paths traverse the network in parallel, further restricting embolism spread at centrally located constriction points. The combined effect of these changes increases p88 as the xylem strand becomes elongated or increasingly medullated (Fig. 2, B and C). Each incremental increase in complexity thus theoretically yields a marginal improvement in survivorship under drought. We found up to a 2-MPa difference in p88 between the two extreme stele shapes, leading to a potential doubling of the mortality threshold in some species (fig. S4, B and C).

Observed Paleozoic and extant pteridophyte xylem strands span nearly the full range of both network traits, with only the most-resistant ideal end points unoccupied (Fig. 2D and figs. S6 to S8). The least-drought-resistant trait combinations are found exclusively in Paleozoic specimens [one-way analysis of variance (ANOVA) indicates the difference from extant plants with $N = 60$, degrees of freedom = 59, $F > 52$, and $p < 0.001$ on both metrics and simulated p88], which corresponds with previously established traits that reduced their drought resistance (*21*, *22*). Extant pteridophytes range down to fewer than three

[1]Institute of Botany, Czech Academy of Sciences, Průhonice, Czechia. [2]Department of Biology, Bates College, Lewiston, ME, USA. [3]Yale School of the Environment, New Haven, CT, USA. [4]New York Botanical Garden, Bronx, NY, USA. [5]School of Forest Resources, University of Maine, Orono, ME, USA. [6]Department of Environmental Studies, Haverford College, Haverford, PA, USA.
*Corresponding author. Email: martin.bouda@ibot.cas.cz (M.B.); craig.brodersen@yale.edu (C.R.B.)

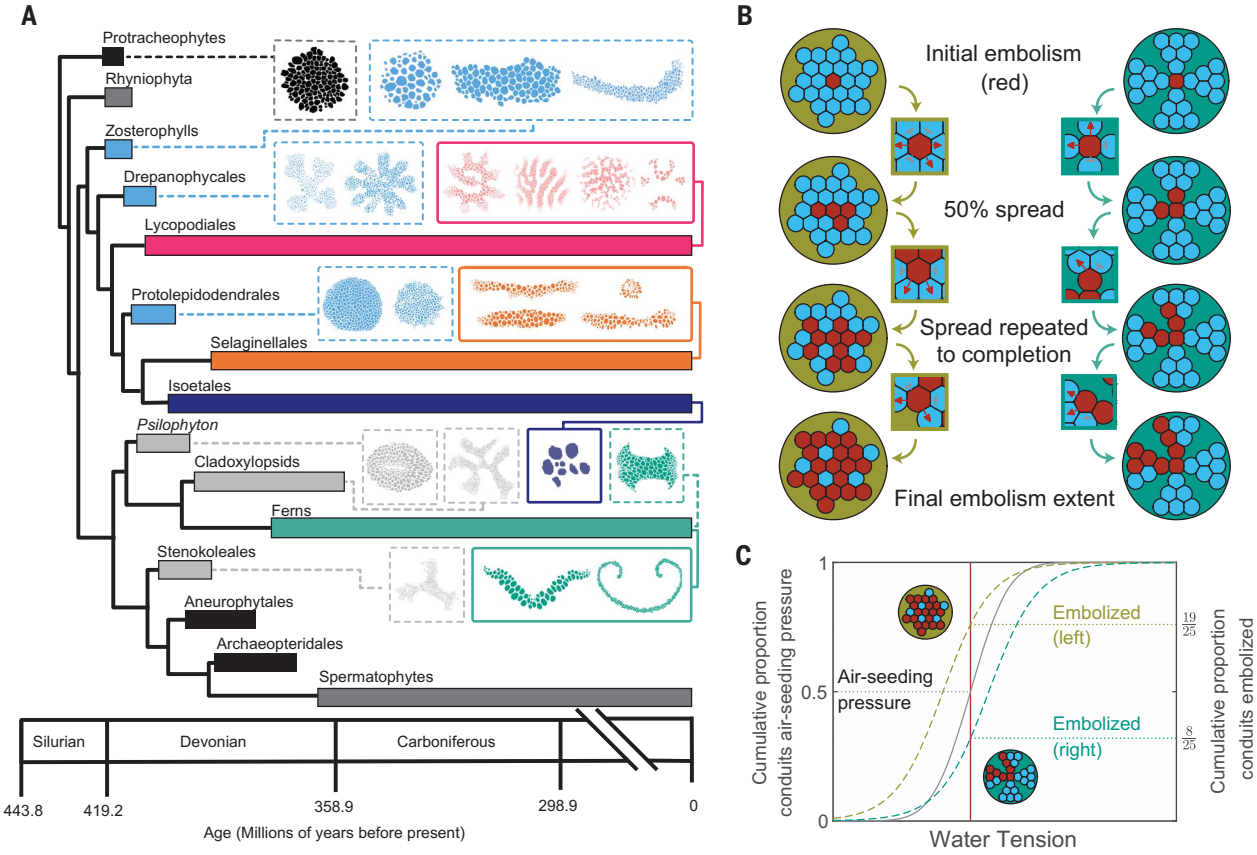


Fig. 1. Xylem strand shapes that are associated with increasing drought tolerance. (**A**) Evolutionary relationships between the tracheophytes and their stratigraphic ranges. Representative xylem strand shapes of Paleozoic (boxes with dashed borders) and extant (boxes with solid borders) vascular plants show a trajectory from the simple terete haplostele to the more-elongated or deeply medullated xylem strand shapes that are observable in more recent taxa. (**B**) Conceptual model of the spread of embolism (red) from a single conduit in two different xylem strand shapes. In both example steles, the scenario represents a worst-case initial embolism position and maximum likelihood spread when water tension corresponds to the median conduit ASP: Embolism spreads from each cell to about half of its neighbors. Our formal simulations randomized initial position and ASP of individual conduit walls over N = 5000 replicates per tension level. (**C**) Corresponding vulnerability curves show that xylem strand shape alone can increase drought resistance, regardless of initial embolus position in the network.

neighbors per conduit and tend to keep below a lnPC of 0, with notable exceptions such as the long, narrow, corrugated xylem strand of *Dicksonia antarctica* or the apparent protostele with deep parenchymal intrusions of *Lygodium microphyllum* (figs. S3 and S9). Although the two metrics are partially correlated, these results suggest that both local reduction in connectivity and overall shape change can yield increasingly drought-resistant xylem.

Among extant species, we found vulnerable topologies only in species limited to ever-moist environments (Fig. 2D and table S1 for habitat information). Whereas we did find mesic species with both relatively vulnerable (e.g., *Selaginella selaginoides*) and resistant xylem topologies (*Osmunda regalis*), we only found xeric species with resistant ones (e.g., *Astrolepis sinuata*, *Cheilanthes distans*). Because no countervailing selection pressure is known to favor less-medullated steles in moister conditions, evolutionary legacies (*23*) may confer more-drought-resistant topologies on extant mesic species. A simple correlation is likely to be further confounded by the fact that drought resistance is a complex plant property that involves many traits operating at multiple spatial and temporal scales, including pit membrane properties, stomatal behavior, leaf structure, drought deciduousness, root hydraulic architecture, and their physiological coordination (*24*). Moreover, available habitat information fails to account for possible confounding microclimate effects. Unaccountable microsite differences in wetness would further undermine a direct correlation. Nevertheless, across plant taxa, habits, and habitats and across a wide range in drought tolerance, mortality is known to occur after stomatal closure only once embolism begins to spread in the xylem (*1*). Resistance to embolism spread is thus a key trait that provides a marginal increase in survivorship among vascular plants on the verge of drought mortality. To the extent that xylem network topology contributes to embolism resistance, drought was likely a major driver of the changes in stelar morphology that have been observed over evolutionary history. Because our Paleozoic species are thought to have occupied various wetland habitats (*25*), our observations are fully consistent with the idea of selection for resistant topologies in drier environments at this macroevolutionary scale.

Network trait analysis allows us to restate the Bower-Wardlaw hypothesis (*7*, *8*) as quantitative predictions on how conduit network topology will diverge from the terete arrangement as xylem strands grow in the number of conduits, owing to a selection pressure by hydraulic failure. As conduits are added to a xylem strand that maintains a terete shape, vulnerability increases. By contrast, straps or lobed shapes yield more-resistant topologies as they grow if the independent lobes remain sufficiently narrow (Fig. 3A). Lineages that increase xylem strand size while maintaining the same ASP distribution should thus be

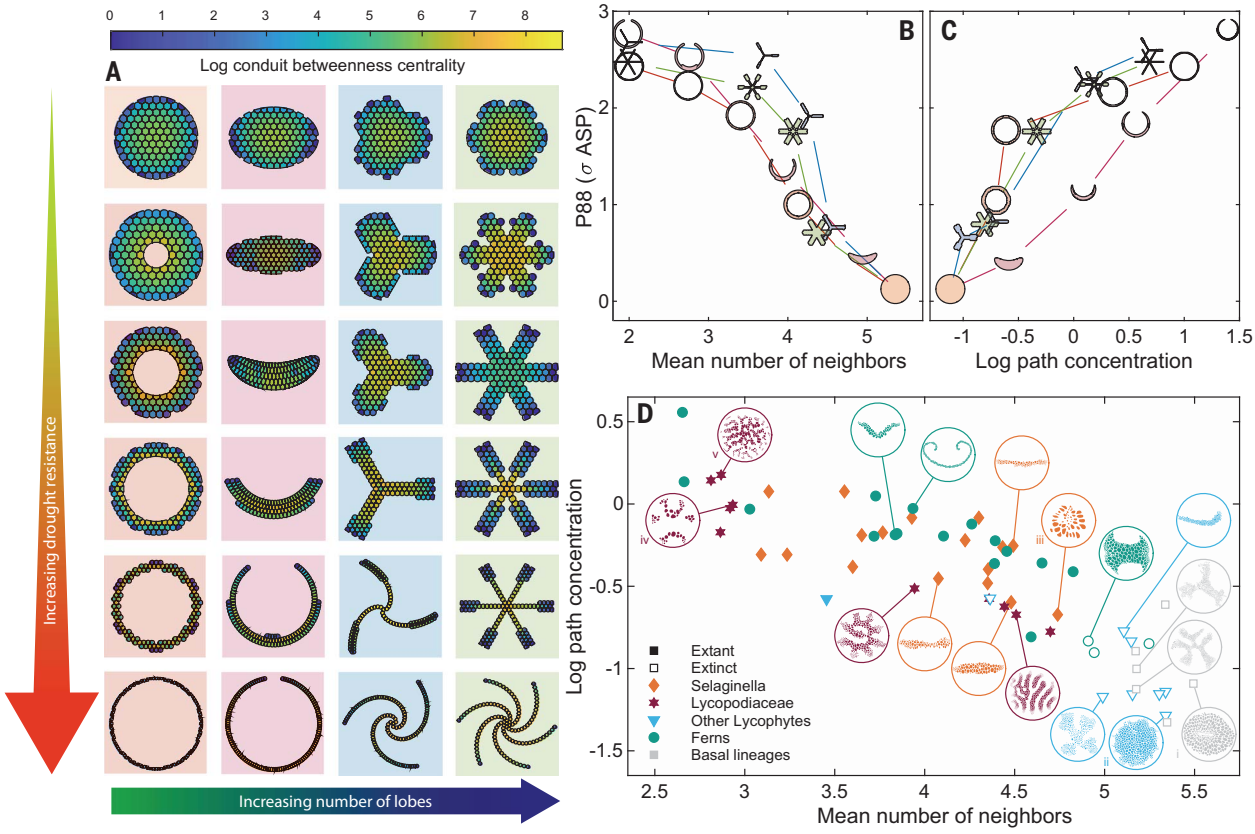


Fig. 2. Theoretical drought-resistance effects of xylem strand shape. (**A**) Increasingly resistant shapes in annular, elliptical, three-lobed, and six-lobed series used in drought simulations. (**B**) p88 versus mean number of neighbors per conduit. (**C**) p88 versus the lnPC. (**D**) Placement of empirical conduit networks of 60 species in space defined by the number of neighbors and lnPC (see table S1 for full list). The insets indicate stele shapes of selected species; species i to v are noted in the main text.

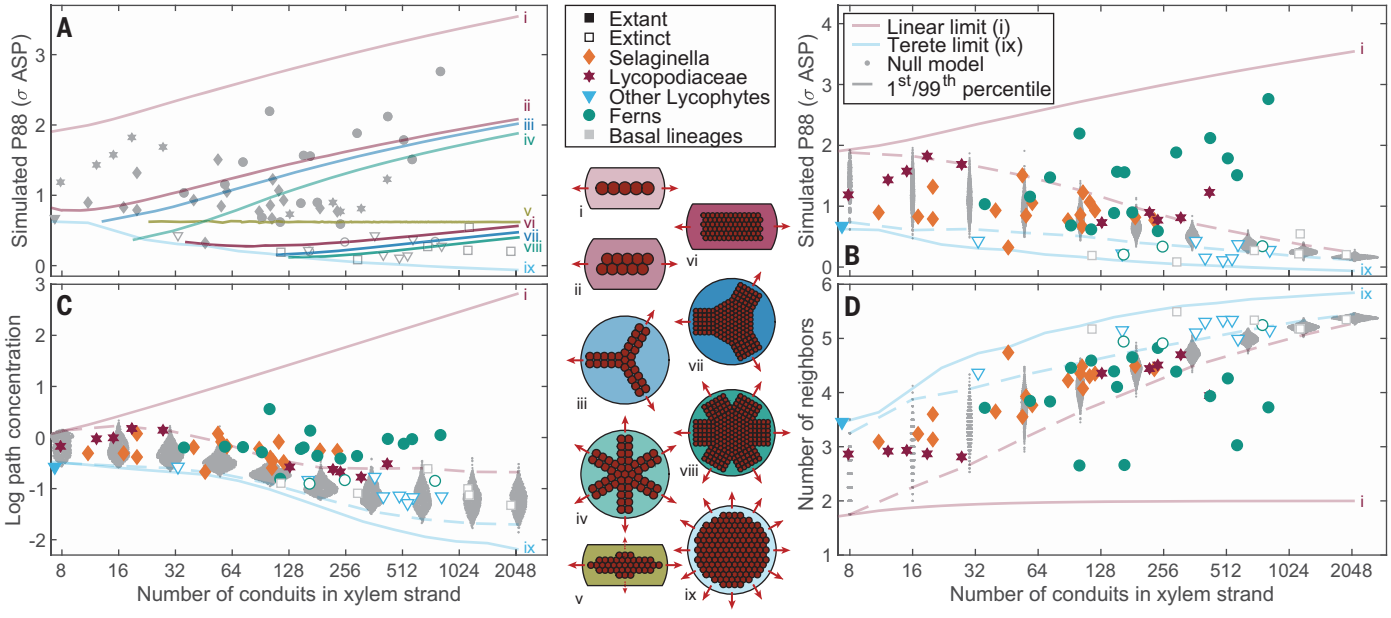


Fig. 3. Xylem strand size, conduit network topology, and drought resistance. (**A** to **D**) Trajectories (A) of the mortality threshold with increasing number of conduits in selected shapes (i to ix). The simulated p88 (B), lnPC (C), and number of neighbors per conduit (D) versus number of conduits in a xylem strand for N = 60 species, overlain on null model distributions of the network traits and their upper and lower bounds.

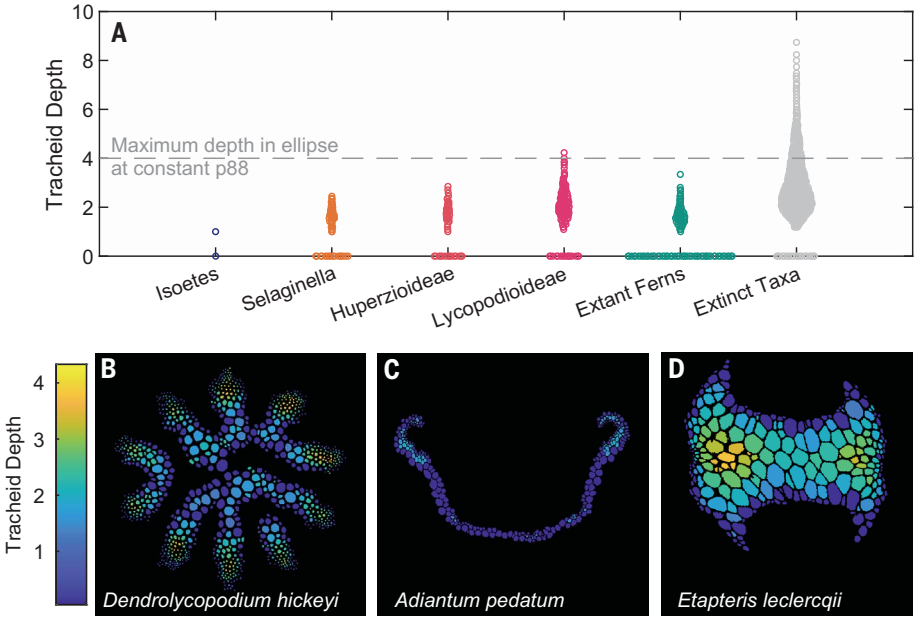


Fig. 4. Xylem strand width in extinct early and extant seedless vascular plants. (**A**) Swarm chart of conduit depth values (number of conduits to strand edge) of individual conduits [distance to strand edge (*20*)] by lineage. (**B** to **D**) Heatmaps of conduit depth in selected extant [(B) and (C)] and fossil (D) species.

selected to diverge from the terete arrangement sufficiently to maintain low vulnerability.

Our observations agree with this prediction. The simulated p88, path concentration, and the number of conduit neighbors are all increasingly divergent from the terete curve in extant species with larger xylem strands (Fig. 3, B to D). Compared with a null model of possible xylem strand shapes at any given size, extant plants have an increasing tendency to appear as strong outliers on path concentration and embolism-spread resistance with an increasing number of conduits [$p < 0.01$ (*20*)]. Paleozoic specimens not only present a much less favorable combination of traits but their divergence from the terete shape also begins at larger sizes and proceeds more gradually than for the extant ones.

Restricting the width of a growing xylem strand is likely a key means of maintaining a drought-resistant topology. Allowing growth from a small terete haplostele to proceed separately along a major and minor axis results almost entirely in elongation along the major axis if we require a constant mortality threshold to be maintained (curve v in Fig. 3A). The shape remains less than eight conduits wide even as it grows beyond 2000 conduits overall. By expressing xylem strand width as the maximum depth of individual conduits (*20*), a survey of pteridophyte steles (Fig. 4 and fig. S10) shows that extant xylem strands are restricted to values less than four (Fig. 4, A to C), which is true for many (Fig. 4D), but certainly not all, fossil steles. This apparent threshold corresponds to the upper bound on conduit depth that is found in the drought-constrained growth simulations, raising the intriguing possibility that selection has favored xylem strands to stay within such a width limit.

The fossil record abounds with examples of parallel evolution, converging on functionally analogous configurations of increased medullation or outright dissection. These include independent originations of deeply lobed actinosteles (e.g., *Asteroxylon*, *Asteropteris*, *Tristichia longii*), multiple separate pathways to the emergence of a pith [e.g., medullated protosteles in Osmundaceae and Lepidodendrales (*6*), concentric siphonosteles common in Polypodiidae ferns, and peripheral placement of discrete vascular bundles in eusteles of the Spermatophyta], or repeated separation of the xylem into multiple strands (polystely in *Selaginella* spp., well-developed braided plectosteles such as those in *Xenocladia medullosina*, separate xylem strands in dictyosteles of many ferns from the Polypodiales order, dissected steles of Cladoxylopsida, and eusteles of the Spermatophyta, including independent origins of anomalous examples in extinct lineages such as *Medullosa* and *Pentoxylon*). These parallel increases in vascular complexity likely confer greater drought survivorship based on our analysis, convergent in function even when divergent in form. Early stelar diversification occurred during a brief period of the Devonian that was marked by decreasing atmospheric CO_2 concentration, when selective pressure on hydraulic safety traits needed to sustain photosynthesis was likely increased (*26*).

Although developmental links between plant stature, branching, and stelar morphology (*3*) are often interpreted as evidence that stele shape responded to innovations in plant body construction, this view suffers from a cause-and-effect dilemma. If stelar morphology is independently subject to selection by drought, the direction of causality may well be reversed. If, as we suggest, selection by drought is a stronger determinant of vascular organization than previously appreciated, its role in the development of branching habits and the eventual emergence of arborescence based on secondary growth also needs to be reexamined.

The proposed selection pressure by drought is theoretically sound given the established linkage between hydraulic failure, tissue death (*16*), and plant mortality (*17*) and is consistent with evidence for selection of increasingly resistant or dissected vascular systems in drying climates (*21*, *27*, *28*). As such, drought resistance provides the mechanistic basis for a substantial selective pressure that drives increased complexity of stelar morphology by simple developmental modifications of conduit network topology. How this process played out in the earliest terrestrial vascular plants is not fully understood, although anatomical evidence points toward selection for drought-resistant tracheids among the lineages leading to the lycophytes, ferns, and seed plants (*10*, *14*), along with other key drought tolerance traits. Applying our understanding of plant hydraulics (*19*, *21*, *22*) to evidence from the fossil record and extant pteridophytes indicates that the observed diversification in stelar morphology decreased xylem network vulnerability. Innovations to vascular organization, in tandem with anatomical adaptations, likely played an important role in the Devonian radiation (*29*, *30*) and the subsequent expansion of terrestrial ecosystems, facilitating the colonization of drier habitats and the development of taller and increasingly branched growth forms.

REFERENCES AND NOTES

1. T. J. Brodribb, J. Powers, H. Cochard, B. Choat, *Science* **368**, 261–266 (2020).
2. J. A. Raven, *Bot. J. Linn. Soc.* **88**, 105–126 (1984).
3. P. Kenrick, P. R. Crane, *Nature* **389**, 33–39 (1997).
4. J. C. McElwain, *Annu. Rev. Plant Biol.* **69**, 761–787 (2018).
5. R. Schmid, *Bot. Rev.* **48**, 817–931 (1982).
6. C. B. Beck, R. Schmid, G. W. Rothwell, *Bot. Rev.* **48**, 691–815 (1982).
7. F. O. Bower, *Proc. R. Soc. Edinb.* **41**, 1–25 (1922).
8. C. W. Wardlaw, *Trans. R. Soc. Edinb.* **53**, 503–532 (1924).
9. K. J. Niklas, *Ann. Bot.* **42**, 33–39 (1978).
10. K. J. Niklas, *Evolution* **39**, 1110–1122 (1985).
11. D. C. Wight, *Paleobiology* **13**, 208–214 (1987).
12. A. M. F. Tomescu, *Biol. Rev. Camb. Philos. Soc.* **96**, 1263–1283 (2021).
13. J. S. Suissa, W. E. Friedman, *Proc. Biol. Sci.* **289**, 20212209 (2022).
14. D. Edwards, C.-S. Li, J. A. Raven, *Bot. J. Linn. Soc.* **150**, 115–130 (2006).
15. M. T. Tyree, M. H. Zimmermann, *Xylem Structure and the Ascent of Sap* (Springer Series in Wood Science, Springer, 2002).

16. T. Brodribb *et al*., *New Phytol.* **232**, 68–79 (2021).
17. M. Mantova, S. Herbette, H. Cochard, J. M. Torres-Ruiz, *Trends Plant Sci.* **27**, 335–345 (2022).
18. A. Mrad, J.-C. Domec, C.-W. Huang, F. Lens, G. Katul, *Plant Cell Environ.* **41**, 2718–2730 (2018).
19. J. Wason *et al*., *Plant Physiol.* **186**, 373–387 (2021).
20. Materials and methods are available as supplementary materials.
21. J. P. Wilson, A. H. Knoll, *Paleobiology* **36**, 335–355 (2010).
22. J. P. Wilson, *Paleontol. Soc. Papers* **19**, 175–194 (2013).
23. J. Cavender-Bares, D. D. Ackerly, S. E. Hobbie, P. A. Townsend, *Annu. Rev. Ecol. Evol. Syst.* **47**, 433–462 (2016).
24. B. Choat *et al*., *Nature* **558**, 531–539 (2018).
25. S. F. Greb, W. A. DiMichele, R. A. Gastaldo, in *Wetlands Through Time*, vol. 399 (Geological Society of America, 2006).
26. B. Chen *et al*., *Earth Sci. Rev.* **222**, 103814 (2021).
27. M. Larter *et al*., *New Phytol.* **215**, 97–112 (2017).
28. H. J. Schenk *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **105**, 11248–11253 (2008).
29. A. H. Knoll, K. J. Niklas, B. H. Tiffney, *Science* **206**, 1400–1402 (1979).
30. K. J. Niklas, B. H. Tiffney, A. H. Knoll, *Nature* **303**, 614–616 (1983).
31. M. Bouda *et al*., Supplementary raw data and code for the manuscript "Drought resistance as a primary driver of stelar evolution in early vascular plants". OSF (2022); https://doi.org/10.17605/OSF.IO/9AMBV.

ACKNOWLEDGMENTS

We thank P. Crane for early discussions of this work. We thank B. Ambrose (New York Botanical Garden), C. Jones and M. Opel (University of Connecticut), and K. Kim (Yale Marsh Botanical Garden) for access to plant material used in this study. We further thank J. Galtier, B. Meyer-Berthaud, and A.-L. Decombeix for images and discussions of Devonian and Carboniferous vascular plants. A. Gandolfo of the Cornell University Plant Anatomy Collection and S. Hu of the Yale Peabody Paleobotanical collection provided access to images and specimens of fossilized plants. We also thank M. Duguid and L. Green of Yale for their assistance in locating several of the ferns and lycophytes used in this study at the Yale Myers Forest. The *S. selaginoides* accession was collected by S. Lavergne. We thank three anonymous reviewers for their insightful comments that greatly improved previous drafts of this manuscript. **Funding:** J.W.W. was partially supported by the US Department of Agriculture, National Institute of Food and Agriculture, McIntire Stennis Project number ME0-42121 through the Maine Agricultural and Forest Experiment Station. B.A.H. was supported by the Phillips Fellowship, Bates College. M.B. and C.R.B. were supported by the Newman Family Plant Research Fund. M.B. was partially supported by long-term research development project no. RVO 67985939 of the Czech Academy of Sciences. Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA CZ LM2018140 and e-INFRA CZ ID:90140) supported by the Ministry of Education, Youth and Sports of the Czech Republic. **Author contributions:** Conceptualization: M.B., C.R.B., J.W.W., J.P.W.; Methodology: M.B., C.R.B., B.A.H.; Software: M.B.; Formal analysis: M.B.; Investigation: M.B., C.R.B., K.A.P., B.A.H., J.P.W.; Resources: M.B., C.R.B., J.P.W.; Visualization: M.B., C.R.B.; Funding acquisition: C.R.B.; Writing – original draft: M.B., C.R.B.; Writing – review and editing: M.B., C.R.B., J.W.W., K.A.P., B.A.H., J.P.W. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All raw images and data used in the analysis, including all data reported in figures here, as well as all original code required to perform the analysis in this work are available online at OSF (*31*). **License information:** Copyright © 2022 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.add2910
Materials and Methods
Figs. S1 to S10
Table S1
References (*32–100*)
Movies S1 to S4

View/request a protocol for this paper from *Bio-protocol*.

Submitted 2 June 2022; accepted 13 October 2022
10.1126/science.add2910

NEUTRON STARS

Polarized x-rays from a magnetar

Roberto Taverna[1]*, Roberto Turolla[1,2], Fabio Muleri[3], Jeremy Heyl[4], Silvia Zane[2], Luca Baldini[5,6], Denis González-Caniulef[4], Matteo Bachetti[7], John Rankin[3], Ilaria Caiazzo[8], Niccolò Di Lalla[9], Victor Doroshenko[10], Manel Errando[11], Ephraim Gau[11], Demet Kırmızıbayrak[4], Henric Krawczynski[11], Michela Negro[12,13,14], Mason Ng[15], Nicola Omodei[9], Andrea Possenti[7], Toru Tamagawa[16,17,18], Keisuke Uchiyama[17,18], Martin C. Weisskopf[19], Ivan Agudo[20], Lucio A. Antonelli[21,22], Wayne H. Baumgartner[19], Ronaldo Bellazzini[6], Stefano Bianchi[23], Stephen D. Bongiorno[19], Raffaella Bonino[24,25], Alessandro Brez[6], Niccolò Bucciantini[26,27,28], Fiamma Capitanio[3], Simone Castellano[6], Elisabetta Cavazzuti[29], Stefano Ciprini[22,30], Enrico Costa[3], Alessandra De Rosa[3], Ettore Del Monte[3], Laura Di Gesu[29], Alessandro Di Marco[3], Immacolata Donnarumma[29], Michal Dov iak[31], Steven R. Ehlert[19], Teruaki Enoto[16], Yuri Evangelista[3], Sergio Fabiani[3], Riccardo Ferrazzoli[3], Javier A. Garcia[32], Shuichi Gunji[33], Kiyoshi Hayashida[34]†, Wataru Iwakiri[35], Svetlana G. Jorstad[36,37], Vladimir Karas[31], Takao Kitaguchi[16], Jeffery J. Kolodziejczak[19], Fabio La Monaca[3], Luca Latronico[24], Ioannis Liodakis[38], Simone Maldera[24], Alberto Manfreda[6], Frédéric Marin[39], Andrea Marinucci[29], Alan P. Marscher[36], Herman L. Marshall[15], Giorgio Matt[23], Ikuyuki Mitsuishi[40], Tsunefumi Mizuno[41], Stephen C.-Y. Ng[42], Stephen L. O'Dell[19], Chiara Oppedisano[24], Alessandro Papitto[21], George G. Pavlov[43], Abel L. Peirson[9], Matteo Perri[22,21], Melissa Pesce-Rollins[6], Maura Pilia[7], Juri Poutanen[44,45], Simonetta Puccetti[22], Brian D. Ramsey[19], Ajay Ratheesh[3], Roger W. Romani[9], Carmelo Sgrò[6], Patrick Slane[46], Paolo Soffitta[3], Gloria Spandre[6], Fabrizio Tavecchio[47], Yuzuru Tawara[40], Allyn F. Tennant[19], Nicholas E. Thomas[19], Francesco Tombesi[48], Alessio Trois[7], Sergey S. Tsygankov[44,45], Jacco Vink[49], Kinwah Wu[2], Fei Xie[50]

Magnetars are neutron stars with ultrastrong magnetic fields, which can be observed in x-rays. Polarization measurements could provide information on their magnetic fields and surface properties. We observed polarized x-rays from the magnetar 4U 0142+61 using the Imaging X-ray Polarimetry Explorer and found a linear polarization degree of 13.5 ± 0.8% averaged over the 2– to 8–kilo–electron volt band. The polarization changes with energy: The degree is 15.0 ± 1.0% at 2 to 4 kilo–electron volts, drops below the instrumental sensitivity ~4 to 5 kilo–electron volts, and rises to 35.2 ± 7.1% at 5.5 to 8 kilo–electron volts. The polarization angle also changes by 90° at ~4 to 5 kilo–electron volts. These results are consistent with a model in which thermal radiation from the magnetar surface is reprocessed by scattering off charged particles in the magnetosphere.

Isolated neutron stars (NSs) with extremely strong magnetic fields are referred to as magnetars (*1*). There are ~30 confirmed magnetars known (*2*), many of which are detectable only during periods of enhanced activity. Magnetar emission is powered by the magnetic field, producing bursts of hard (≈10 to 100 keV) x-rays, with luminosity $L \approx 10^{38}$ to 10^{47} erg s^{-1} and duration ≈ 0.1 to 100 s. Magnetars also exhibit persistent x-ray emission at $L \approx 10^{33}$ to 10^{35} erg s^{-1}, which is pulsed at spin frequencies $f \approx 0.1$ to 10 Hz with spin-down rates $\dot{f} \approx -(10^{-16}\text{to}10^{-8})\text{Hzs}^{-1}$. These properties indicate high magnetic fields $B \lesssim 10^{15}$G, assuming a standard spin-down model (*3*). The 0.5- to 10-keV spectrum of magnetars consists of a blackbody (BB) component (with $kT \sim 0.1$ to 1 keV, where T is the temperature and k is the Boltzmann constant) and a power-law (PL) component with photon index $\Gamma \sim 2$ to 4; the PL dominates above ~4 to 5 keV (*2*, *3*). Some sources exhibit a second BB component instead of the PL. Many magnetars are detected in x-rays up to ≈200 keV, at which the spectrum is also dominated by the PL component.

The magnetic field surrounding magnetars is expected to differ from a pure dipole, with a non-negligible toroidal component that twists the field lines. Because charged particles flow along closed magnetic field lines, as required to sustain the field, the region threaded by the magnetic field (the magnetosphere) becomes optically thick to Compton scattering at the cyclotron resonance frequency [resonant Compton scattering (RCS)] (*4*). The BB spectral component is expected to be emitted by (multiple regions on) the cooling surface of the NS, whereas the PL originates from the reprocessing of thermal photons through resonant up-scattering in the magnetosphere (*3*).

Magnetar x-ray persistent emission is expected to be linearly polarized in two orthogonal modes, referred to as ordinary (O) and extraordinary (X), with the polarization vector either parallel or perpendicular to the plane formed by the photon propagation direction and the (local) magnetic field (*5*). The expected polarization degree of the emitted radiation strongly depends on the physical state of the NS external layers. If radiation comes from the bare, condensed surface, the polarization is expected to be $\lesssim 10\%$, but a magnetized atmosphere can produce polarization $\lesssim 80\%$ (*6*–*8*). The polarization of outgoing photons is then modified by RCS, which leads to a polarization degree $\lesssim 30\%$ in the X mode for the

[1]Department of Physics and Astronomy, University of Padova, I-35131 Padova, Italy. [2]Mullard Space Science Laboratory, University College London, Holmbury St Mary Dorking RH5 6NT, UK. [3]Istituto di Astrofisica e Planetologia Spaziali, Istituto Nazionale di Astrofisica (INAF), I-00133 Roma, Italy. [4]Department of Physics and Astronomy, University of British Columbia, Vancouver, BC V6T 1Z1, Canada. [5]Dipartimento di Fisica Enrico Fermi, Università di Pisa, I-56127 Pisa, Italy. [6]Istituto Nazionale di Fisica Nucleare (INFN) Sezione di Pisa, I-56127 Pisa, Italy. [7]Osservatorio Astronomico di Cagliari, INAF, I-09047 Selargius, Italy. [8]Theoretical AstroPhysics Including Relativity and Cosmology, Caltech, Pasadena, CA 91125, USA. [9]Deparment of Physics and Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, Stanford, CA 94305, USA. [10]Institut für Astronomie und Astrophysik, Universität Tübingen, 72076 Tübingen, Germany. [11]Physics Department and McDonnell Center for the Space Sciences, Washington University, St. Louis, MO 63130, USA. [12]Center for Space Sciences and Technology, University of Maryland, Baltimore, MD 21250, USA. [13]NASA Goddard Space Flight Center (GSFC), Greenbelt, MD 20771, USA. [14]Center for Research and Exploration in Space Science and Technology, NASA/GSFC, Greenbelt, MD 20771, USA. [15]Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [16]RIKEN Cluster for Pioneering Research, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan. [17]RIKEN Nishina Center, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan. [18]Department of Physics, Tokyo University of Science, 1-3 Kagurazaka, Shinjuku, Tokyo 162-8601, Japan. [19]NASA Marshall Space Flight Center (MSFC), Huntsville, AL 35812, USA. [20]Instituto de Astrofísica de Andalucía, 18008 Granada, Spain. [21]Osservatorio Astronomico di Roma, INAF, 00040 Monte Porzio Catone, Italy. [22]Space Science Data Center (SSDC), Agenzia Spaziale Italiana (ASI), 00133 Roma, Italy. [23]Dipartimento di Matematica e Fisica, Università degli Studi Roma Tre, 00146 Roma, Italy. [24]INFN Sezione di Torino, 10125 Torino, Italy. [25]Dipartimento di Fisica, Università degli Studi di Torino, 10125 Torino, Italy. [26]Osservatorio Astrofisico di Arcetri, INAF, 50125 Firenze, Italy. [27]Dipartimento di Fisica e Astronomia, Università degli Studi di Firenze, 50019 Sesto Fiorentino, Italy. [28]INFN Sezione di Firenze, 50019 Sesto Fiorentino, Italy. [29]ASI, 00133 Roma, Italy. [30]INFN Sezione di Roma Tor Vergata, 00133 Roma, Italy. [31]Astronomical Institute of the Czech Academy of Sciences, 14100 Praha 4, Czech Republic. [32]Cahill Center for Astronomy and Astrophysics, Caltech, Pasadena, CA 91125, USA. [33]Department of Physics, Yamagata University, 1-4-12 Kojirakawa-machi, Yamagata-shi 990-8560, Japan. [34]Department of Earth and Space Science, Osaka University, 1-1 Yamadaoka, Suita, Osaka 565-0871, Japan. [35]Department of Physics, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan. [36]Institute for Astrophysical Research, Boston University, Boston, MA 02215, USA. [37]Laboratory of Observational Astrophysics, St. Petersburg University, St. Petersburg 199034, Russia. [38]Finnish Centre for Astronomy with the European Southern Observatory, University of Turku, 20014 Turku, Finland. [39]Observatoire Astronomique de Strasbourg, Université de Strasbourg, 67000 Strasbourg, France. [40]Division of Particle and Astrophysical Science, Graduate School of Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8602, Japan. [41]Hiroshima Astrophysical Science Center, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8526, Japan. [42]Department of Physics, The University of Hong Kong, Pokfulam, Hong Kong. [43]Department of Astronomy and Astrophysics, Pennsylvania State University, University Park, PA 16801, USA. [44]Department of Physics and Astronomy, University of Turku, 20014 Turku, Finland. [45]Space Research Institute of the Russian Academy of Sciences, Moscow 117997, Russia. [46]Center for Astrophysics, Harvard & Smithsonian, Cambridge, MA 02138, USA. [47]Osservatorio Astronomico di Brera, INAF, 23807 Merate, Italy. [48]Dipartimento di Fisica, Università degli Studi di Roma Tor Vergata, 00133 Roma, Italy. [49]Anton Pannekoek Institute for Astronomy, University of Amsterdam, 1098 XH Amsterdam, Netherlands. [50]Guangxi Key Laboratory for Relativistic Astrophysics, School of Physical Science and Technology, Guangxi University, Nanning 530004, China.
*Corresponding author. Email: taverna@pd.infn.it
†Deceased.

PL component, independent of the initial polarization state of the thermal photons (*7*–*9*).

Because NSs cannot be spatially resolved by observations, the contributions from regions with different magnetic field orientations (and therefore with different emitted polarization orientations) are blended together, which reduces the observed polarization (*10*, *11*). However, if the magnetic field is strong enough (*5*), it forces the photon polarization vectors to follow the magnetic field direction, which results in an observed polarization almost unchanged from that at the emission (*10*, *11*).

The magnetar 4U 0142+61 (coordinates right ascension $01^{h}\ 46^{m}\ 22^{s}.41$, declination 61° 45′ 03″.2, J2000 equinox) has a persistent (lightly variable) x-ray flux of ~6×10^{-11} erg s^{-1} cm^{-2} in the 2- to 10-keV range, a spin frequency of f = 0.12 Hz, and a frequency derivative of $\dot{f} = -2.6 \times 10^{-14}$Hzs^{-1}. This implies a spin-down (equatorial) magnetic field of $B \sim 1.3 \times 10^{14}$ G (*2*, *12*). It is visible at infrared and optical wavelengths (*13*), but no (pulsed) radio emission has been detected.

We observed 4U 0142+61 with the Imaging X-ray Polarimetry Explorer (IXPE) (*14*) between 31 January 2022 and 27 February 2022 for a total on-source time of 840 ks. IXPE provides imaging polarimetry over a nominal energy band of 2 to 8 keV. The data were extracted and processed according to standard procedures (*15*). Pulsations were detected (fig. S3) at f = 0.115079336 ± 6×10^{-9} Hz with $\dot{f} = -(2.1 \pm 0.7) \times 10^{-14}$Hzs^{-1} (at MJD 59624.050547, where MJD is the modified Julian date); uncertainties are 68.3% confidence. These values are consistent with previous measurements, within the uncertainties (*12*). We performed a spectral analysis using the software package XSPEC (*16*), version 12.12.1. The data are not consistent with a single-component model, so we considered several two-component models (*15*). In all models, we fixed the value of the foreground interstellar column density to 0.57×10^{22} cm^{-2} (*17*); it cannot be constrained by the IXPE data because of insufficient sensitivity below 2 keV. Our best-fitting parameters for a BB + PL model (table S2) are consistent with previous measurements (*17*, *18*).

Polarization was measured by extracting the (calibrated) Stokes parameters I, Q, and U from each photon, which were collected by the three independent IXPE detector units (DUs). After subtracting the sky background, the contributions of each DU were combined to account for the 120° offset between the DUs. Figure 1 shows the phase-averaged, normalized Stokes parameters (Q/I and U/I) in the 2- to 8-keV energy range for the individual DUs and the combined data. The phase-averaged, energy-integrated values are Q/I = 0.013 ± 0.008 and U/I = 0.120 ± 0.008, which implies a polarization degree $PD \equiv \sqrt{Q^2 + U^2}/I$ of 13.5 ± 0.8% and a polarization angle $PA \equiv \arctan(U/Q)/2$ of +48.5° ± 1.6°, with positive values being east of (local celestial) north; uncertainties are 1σ. We derived these values using two different methods and found consistent results (*15*). We determined that the minimum detectable polarization at 99% confidence level (MDP_{99}) for our observation is ~2% over the 2- to 8-keV range, so the significance of the nonzero polarization degree is ~17σ.

To investigate whether the PD and PA depend on the photon energy, the data were grouped into five energy bins, which were selected to contain similar numbers of counts in each bin. Figure 2 shows a polar plot of the results. We find that the PD is 15.0 ± 1.0% at low energies (~2 to 4 keV), ~10σ above the MDP_{99} of that bin, which is ~4%. At 4 to 5 keV, the PD is consistent with zero. In the highest-energy bin (5.5 to 8 keV), the PD is 35.2 ± 7.1%, where the MDP_{99} is ~21%. The PA is ~50° at energies below 4 keV and −40° above 5 keV, a swing of 90°.

We also performed a spectropolarimetric analysis by separately convolving the low- and high-energy spectral components with a constant polarization model (POLCONST in XSPEC). This confirms the 90° swing in polarization angle for all the two-component spectral models we considered: BB + BB, BB + PL, and BB + truncated PL (*15*). For the latter model, the derived PD for the two components is within ~1σ of the observed values, with the low-energy BB component being less polarized than the high-energy PL (*15*).

To perform a phase-dependent analysis, we divided the flux into 100 phase bins and used an unbinned maximum likelihood technique (*19*) to determine the PD and PA. Figure 3A shows the resulting pulse profile, which is double peaked, as found in previous observations (*18*). Phase variations are evident in both PD and in PA (Fig. 3, B and C), with

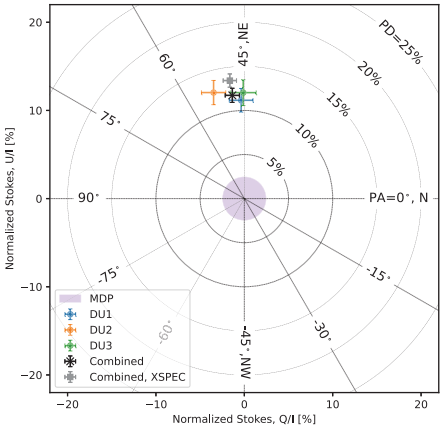


Fig. 1. Normalized, background-subtracted Stokes parameters *Q*/*I* and *U*/*I* for x-ray emission from 4U 0142+61. The values measured from each of the three IXPE DUs (in the 2- to 8-keV range) are marked by green, orange, and blue circles with 1σ error bars, and their combinations obtained using two approaches (*15*) are shown by the black cross and the gray square, respectively. The background circles indicate PD, and the radial lines indicate PA, measured east from north. The purple shaded area shows the detection limit (MDP_{99}) for the combined measurement.

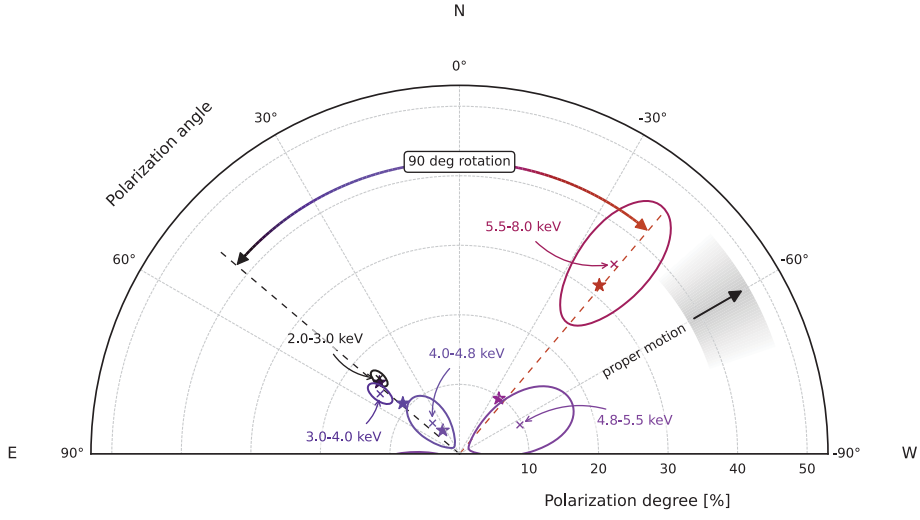


Fig. 2. Polar plot showing the energy dependence of the measured PD and PA. Crosses indicate the measured values in labeled energy bins, and contours enclose the 68.3% confidence level regions obtained with XSPEC (*15*). Stars indicate the corresponding PD and PA calculated using the condensed-surface RCS model. The arc bounded by the two dashed lines shows the change in polarization angle from the lowest (2 to 3 keV, black dashed line) to the highest (5.5 to 8 keV, red dashed line) energy bins. The black arrow and gray shaded area indicate the proper motion direction of the source and its associated uncertainty (*28*).

amplitudes of ~10% and ~30°, respectively. At low energies (2 to 4 keV), we find the main and secondary peaks have higher polarization fraction (~15%) than the phase valley between them (~9%). By contrast, the phase-resolved PA is single peaked. This is consistent with the predictions of pulsar (a different type of NS) models [specifically the rotating-vector model (*20*)], although a strong degeneracy prevents us from determining the NS spin and magnetic axes orientations from the PA data (*15*).

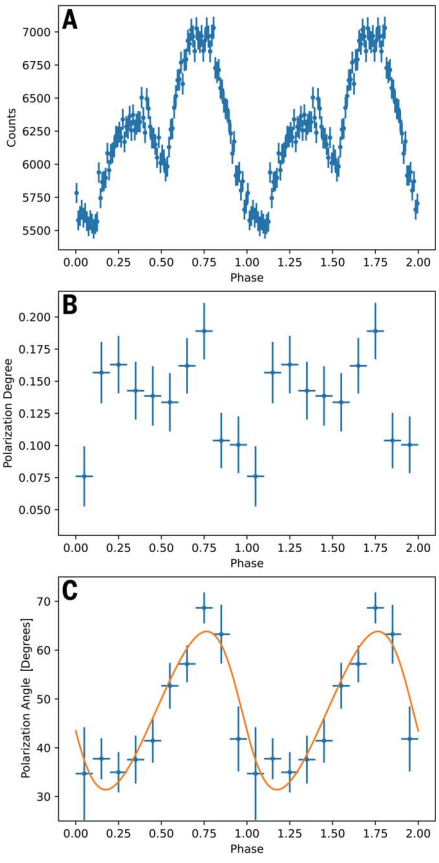


Fig. 3. Phase-dependent x-ray flux and polarization properties. (**A**) Energy-integrated (2 to 8 keV) IXPE counts as a function of spin phase. Error bars are at 1σ confidence level. (**B**) Polarization degree as a function of spin phase. Error bars indicate $\Delta \log L = 1$, where L is the unbinned likelihood (*19*). (**C**) Same as (B), but for the polarization angle. The orange curve shows the best-fitting rotating vector model (*15*).

A phase-resolved spectral analysis of 4U 0142+61 shows no statistically significant dependence of the spectrum on rotational phase (*15*). The BB component is compatible with being constant in phase (fig. S5), which is consistent with previous results (*21*) and previous observations of a low pulsed fraction (~5%) below 3 to 4 keV (*18*).

We considered the IXPE results within a twisted-magnetosphere model (*4*), accounting for the quantum electrodynamical effect of vacuum birefringence (*7*–*9*). The observed polarization behavior as a function of energy—with a minimum PD and a 90° swing of PA at 4 to 5 keV—indicates that the 2- to 8-keV x-ray emission from 4U 0142+61 has two distinct components, polarized in two different normal modes, which correspond to the two components identified in the spectral analysis. In this framework, the low-energy component is produced by thermal emission from the surface of the NS, whereas the high-energy component is produced by photons scattered to higher energies in the magnetosphere (Fig. 4A). The measured polarization fraction at high energies (~35% at 5.5 to 8 keV) is compatible with the theoretical prediction of the RCS model (*7*) and indicates that X-mode photons dominate at high energies; conversely, O-mode photons dominate at low energies.

Theoretical models for magnetar surface emission of soft x-rays predict either (i) a large ($\gtrsim$50%) polarization degree in the X mode if there is a gaseous atmosphere heated from below (*22*) or (ii) a small $\lesssim$10% polarization degree in the O mode if there is a condensed (solid or liquid) surface (*6*–*8*, *23*). The IXPE result below 4 keV is not compatible with the presence of an atmosphere and only marginally compatible with a condensed surface. The latter would be more consistent with the data if the PD could be raised in the model, perhaps by thermal radiation being emitted from only a limited region, not the entire surface (as was assumed in previous calculations). The low pulsed fraction at low energies (*18*) indicates an extended emitting area. Using a numerical code (*7*), we calculated that radiation from a condensed iron surface, emitted from an equatorial belt, produces O-mode photons at low energies (2 to 4 keV) with PD ~15%. Reprocessing by RCS then produces an excess of X-mode photons at higher energies (5.5 to 8 keV) with PD ~35%, whereas the PA changes by 90°. Our calculation does not assume that the reference direction in the plane of the sky (from which the PA is computed) coincides with the projection of the NS spin axis. To match the measured and predicted (absolute) values, an offset is added to the simulated PA (*15*). Figure 2 shows the results of our numerical simulation for a magnetic field strength ~10^{14} G, as measured for 4U 0142+61 (*18*), assuming the emissivity of an iron condensed surface (*23*), in the fixed-ion approximation. A hotter belt close to the magnetic equator appears in NS magnetothermal evolution calculations in both two and three dimensions (*24*, *25*).

We also consider alternative models to explain the IXPE data. Within the RCS paradigm, low-energy O-mode photons could be produced by a gaseous layer with an inverted temperature profile, with a downward flow of energy, as might be produced by external particle bombardment (*26*). In this case, O-mode photons would escape from a deeper (and so hotter) region than in a passively cooling atmosphere and would dominate the outgoing flux.

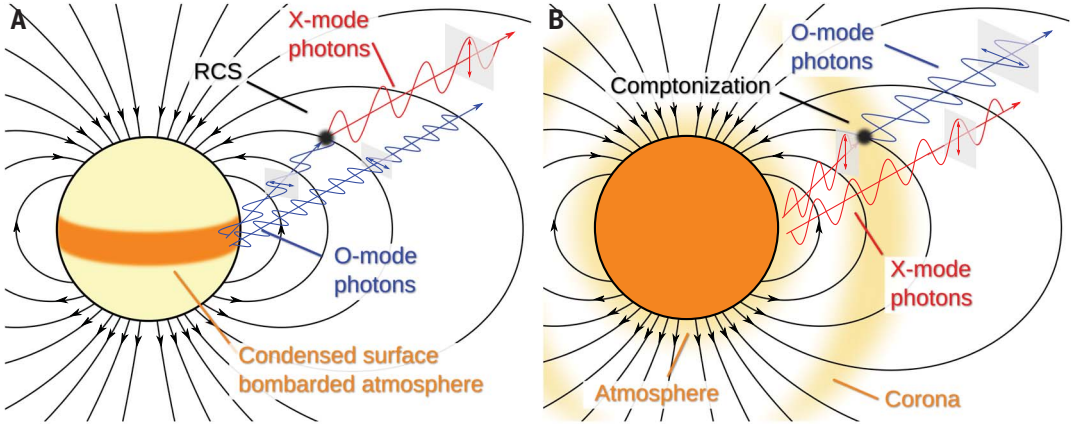


Fig. 4. Schematic illustration of the proposed theoretical scenarios. (**A**) Thermal radiation emitted by an equatorial belt on the condensed surface of the magnetar (or an atmosphere with an inverted temperature gradient), then reprocessed by RCS in the magnetosphere. (**B**) Radiation from the whole surface reprocessed by (unsaturated) thermal Compton scattering in a near-surface atmospheric layer, then additional (saturated) Compton scattering in an extended corona. The dark orange areas on the NS surface indicate the emitting regions. Black lines with arrows indicate the (dipole) magnetic field lines. The gray rectangles along the photon trajectories highlight the polarization plane and the oscillating electric field.

In an alternative scenario, the low-energy emission could be interpreted as polarized in the X mode and the high-energy emission, above 4 to 5 keV, in the O mode. Low-energy, X mode–dominated emission with a low polarization degree (~15%) could originate from an extended region of a condensed iron surface seen few degrees away from the magnetic axis. Radiation from a thin atmosphere or corona in the presence of thermal photons that are undergoing Compton scattering (*8*) could produce the observed polarization at low energies. However, this scenario does not explain how O-mode photons would dominate the emission in the 5- to 8-keV band. Saturated Compton scattering in a thin atmosphere or corona (*8*) or emission from an electron-positron plasma (*27*) could potentially produce O mode–dominated radiation (Fig. 4B), but these models predict a much higher PD than is observed. Emission from a small region of the surface that is covered by an externally illuminated gaseous layer but hot enough to dominate the high-energy band would also produce substantial polarization in the O mode. No detailed modeling of these scenarios is available.

Identifying the mode in which the observed x-ray photons are predominantly polarized would determine the orientation of the magnetar spin axis projected onto the plane of the sky. The phase-averaged PA is 0° (or 90°) for radiation mostly polarized in the O mode (or X mode), taking the reference direction in the plane of the sky to be along the spin axis projection (*10*). If O-mode photons dominate at low energies at which PA ~ 50°, as in the RCS model, the projection of the spin axis would be ~50° east of north. Conversely, if low-energy photons are polarized in the X mode, the spin axis projection would be ~40° west of north. In the latter case, the spin projection would be consistent with the direction of the magnetar proper motion, 60° ± 12° west of north (Fig. 2) (*28*), whereas in the former case, the two would be almost orthogonal. It is unclear which is more appropriate for magnetars. Observations of pulsars (including the Crab Pulsar and Vela Pulsar) show alignment of the spin axis with the proper motion (*29*). However, binary star evolution theory predicts that NSs should be accelerated perpendicular to their spin axis during their formation process (*30*). We are unable to distinguish between these possibilities.

We have detected (linearly) polarized x-ray emission from the magnetar 4U 0142+61. The polarization properties vary with x-ray energy, including a 90° swing of the polarization angle. These observations can be explained by a model of emission from the bare condensed surface of the NS that is reprocessed by RCS in a twisted magnetosphere. Alternative explanations are also possible.

REFERENCES AND NOTES

1. R. C. Duncan, C. Thompson, *Astrophys. J.* **392**, L9 (1992).
2. S. A. Olausen, V. M. Kaspi, *Astrophys. J. Suppl. Ser.* **212**, 6 (2014).
3. V. M. Kaspi, A. M. Beloborodov, *Annu. Rev. Astron. Astrophys.* **55**, 261–301 (2017).
4. C. Thompson, M. Lyutikov, S. R. Kulkarni, *Astrophys. J.* **574**, 332–355 (2002).
5. A. K. Harding, D. Lai, *Rep. Prog. Phys.* **69**, 2631–2708 (2006).
6. D. González Caniulef, S. Zane, R. Taverna, R. Turolla, K. Wu, *Mon. Not. R. Astron. Soc.* **459**, 3585–3595 (2016).
7. R. Taverna, R. Turolla, V. Suleimanov, A. Y. Potekhin, S. Zane, *Mon. Not. R. Astron. Soc.* **492**, 5057–5074 (2020).
8. I. Caiazzo, D. González-Caniulef, J. Heyl, R. Fernández, *Mon. Not. R. Astron. Soc.* **514**, 5024–5034 (2022).
9. R. Fernández, S. W. Davis, *Astrophys. J.* **730**, 131 (2011).
10. R. Taverna *et al.*, *Mon. Not. R. Astron. Soc.* **454**, 3254–3266 (2015).
11. J. S. Heyl, N. J. Shaviv, D. Lloyd, *Mon. Not. R. Astron. Soc.* **342**, 134–144 (2003).
12. R. Dib, V. M. Kaspi, *Astrophys. J.* **784**, 37 (2014).
13. F. Hulleman, M. H. van Kerkwijk, S. R. Kulkarni, *Nature* **408**, 689–692 (2000).
14. M. C. Weisskopf *et al.*, *J. Astron. Telesc. Instrum. Syst.* **8**, 026002 (2022).
15. Materials and methods are available as supplementary materials.
16. K. A. Arnaud, *Astronomical Data Analysis Software and Systems V*, G. H. Jacoby, J. Barnes, Eds. (Astronomical Society of the Pacific Conference Series, vol. 101, 1996), p. 17.
17. P. R. den Hartog *et al.*, *Astron. Astrophys.* **489**, 245–261 (2008).
18. N. Rea *et al.*, *Mon. Not. R. Astron. Soc.* **381**, 293–300 (2007).
19. D. González-Caniulef, I. Caiazzo, J. Heyl, Unbinned likelihood analysis for x-ray polarization. arXiv:2204.00140 [astro-ph.IM] (2022).
20. V. Radhakrishnan, D. J. Cooke, *Astrophys. J. Lett.* **3**, 225 (1969).
21. S. P. Tendulkar *et al.*, *Astrophys. J.* **808**, 32 (2015).
22. A. Y. Potekhin, G. Chabrier, W. C. G. Ho, *Astron. Astrophys.* **572**, A69 (2014).
23. A. Y. Potekhin, V. F. Suleimanov, M. van Adelsberg, K. Werner, *Astron. Astrophys.* **546**, A121 (2012).
24. D. Viganò, "Magnetic fields in neutron stars," thesis, Universidad de Alicante, Alicante, Spain (2013); https://arxiv.org/abs/1310.1243.
25. D. De Grandis *et al.*, *Astrophys. J.* **914**, 118 (2021).
26. D. González-Caniulef, S. Zane, R. Turolla, K. Wu, *Mon. Not. R. Astron. Soc.* **483**, 599–613 (2019).
27. C. Thompson, A. Kostenko, *Astrophys. J.* **904**, 184 (2020).
28. S. P. Tendulkar, P. B. Cameron, S. R. Kulkarni, *Astrophys. J.* **772**, 31 (2013).
29. H. T. Janka, A. Wongwathanarat, M. Kramer, *Astrophys. J.* **926**, 9 (2022).
30. M. Colpi, I. Wasserman, *Astrophys. J.* **581**, 1271–1279 (2002).
31. R. Taverna, robertotaverna/magMC: magMC, version 1.1.1, Zenodo (2022).

ACKNOWLEDGMENTS

We thank three anonymous referees for helpful and constructive comments that improved the paper. This paper is based on observations made by the IXPE, a joint US and Italian mission. The US contribution to the IXPE mission is supported by NASA and led and managed by the MSFC with industry partner Ball Aerospace (contract NNM15AA18C). The Italian contribution is supported by ASI through contract ASI-OHBI-2017-12-I.0, agreements ASI-INAF-2017-12-H0 and ASI-INFN-2017.13-H0 and the SSDC with agreements ASI-INAF-2022-14-HH.0 and ASI-INFN 2021-43-HH.0, and by INAF and INFN. Data products were provided by the IXPE Team (MSFC, SSDC, INAF, and INFN) and distributed with additional software tools by the High-Energy Astrophysics Science Archive Research Center (HEASARC) at NASA GSFC. **Funding:** R.Ta. and R.Tu. acknowledge financial support from the Italian MUR through grant PRIN 2017LJ39LM. J.H., D.G.-C., I.C., and D.K. acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), funding reference number 5007110, and the Canadian Space Agency. D.G.-C. is a Canadian Institute for Theoretical Astrophysical (CITA) National Fellow (grant no. CITA 490888-16). E.G. and H.K. acknowledge NASA support under grants 80NSSC18K0264, 80NSSC22K1291, 80NSSC21K1817, and NNX16AC24G. M.Ne. acknowledges support from NASA under award 80GSFC21M0002. T.T. was supported by grant JSPS KAKENHI JP19H05609. F.Mu., J.R., S.B., E.Co., E.D.M., A.D.M., Y.E., S.F., R.F., F.L.M., G.M., M.Pe., A.R., P.So., and A.T. were supported by ASI and INAF under grants ASI-INAF-2017-12-H0 and ASI-INAF-2022-14-HH.0. L.B., L.L., R.Be., R.Bo., A.B., S.Ca., S.M., A.Mar., C.O., M.P.-R., C.S., and G.S. were supported by ASI and INFN under grants ASI-INFN-2017.13-H0 and ASI-INFN-2021-43-HH.0. **Author contributions:** R.Ta., R.Tu., F.Mu., J.H., S.Z., L.B., J.R., and M.C.W. planned the observing campaign. R.Ta., R.Tu., F.Mu., J.H., S.Z., L.B., J.R., D.G.-C., M.B., I.C., N.D.L., E.G., D.K., H.K., M.Ne., M. Ng, N.O., A.Po., T.T., and K.U. analyzed the data. R.Ta., R.Tu., F.Mu., J.H., S.Z., L.B., J.R., D.G.-C., M.B., I.C., A.Po., T.T., and K.U. modeled the data. R.Ta., R.Tu., F.Mu., J.H., S.Z., L.B., J.R., D.G.-C., M.B., and A.Po. wrote the manuscript. V.D. and M.E. served as internal reviewers. All the other authors contributed to the design and science case of the IXPE mission and to the planning of the observations in this paper. All authors provided input and comments on the manuscript. **Competing interests:** The authors declare there are no competing interests. **Data and materials availability:** The IXPE observation of 4U 0142+61 is available in the HEASARC IXPE Data Archive https://heasarc.gsfc.nasa.gov/docs/ixpe/archive/ under ObsID 01003299. The IXPE software is available at https://github.com/lucabaldini/ixpeobssim and documented at https://ixpeobssim.readthedocs.io. Our measured polarizations are listed in table S2, and the results of our model fitting are listed in tables S1 and S3. The code used for the equatorial belt simulation is available at https://github.com/robertotaverna/magMC and archived on Zenodo (*31*). **License information:** Copyright © 2022 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.add0080
Materials and Methods
Figs. S1 to S10
Tables S1 to S3
References (*32*–*59*)

Submitted 16 May 2022; accepted 18 October 2022
10.1126/science.add0080

BLACK HOLES

Polarized x-rays constrain the disk-jet geometry in the black hole x-ray binary Cygnus X-1

Henric Krawczynski[1]*, Fabio Muleri[2]*, Michal Dovčiak[3]*, Alexandra Veledina[4,5,6]*, Nicole Rodriguez Cavero[1], Jiri Svoboda[3], Adam Ingram[7], Giorgio Matt[8], Javier A. Garcia[9], Vladislav Loktev[4], Michela Negro[10,11,12], Juri Poutanen[4,6], Takao Kitaguchi[13], Jakub Podgorný[3,14,15], John Rankin[2], Wenda Zhang[16], Andrei Berdyugin[4], Svetlana V. Berdyugina[17,18,19], Stefano Bianchi[8], Dmitry Blinov[20,21], Fiamma Capitanio[2], Niccolò Di Lalla[22], Paul Draghis[23], Sergio Fabiani[2], Masato Kagitani[24], Vadim Kravtsov[4], Sebastian Kiehlmann[20,21], Luca Latronico[25], Alexander A. Lutovinov[6], Nikos Mandarakas[20,21], Frédéric Marin[14], Andrea Marinucci[26], Jon M. Miller[23], Tsunefumi Mizuno[27], Sergey V. Molkov[6], Nicola Omodei[22], Pierre-Olivier Petrucci[28], Ajay Ratheesh[2], Takeshi Sakanoi[24], Andrei N. Semena[6], Raphael Skalidis[20,21], Paolo Soffitta[2], Allyn F. Tennant[29], Phillipp Thalhammer[30], Francesco Tombesi[31,32,33], Martin C. Weisskopf[29], Joern Wilms[30], Sixuan Zhang[27], Iván Agudo[34], Lucio A. Antonelli[35,36], Matteo Bachetti[37], Luca Baldini[38,39], Wayne H. Baumgartner[29], Ronaldo Bellazzini[38], Stephen D. Bongiorno[29], Raffaella Bonino[25,40], Alessandro Brez[38], Niccolò Bucciantini[41,42,43], Simone Castellano[38], Elisabetta Cavazzuti[26], Stefano Ciprini[32,36], Enrico Costa[2], Alessandra De Rosa[2], Ettore Del Monte[2], Laura Di Gesu[26], Alessandro Di Marco[2], Immacolata Donnarumma[26], Victor Doroshenko[44,6], Steven R. Ehlert[29], Teruaki Enoto[13], Yuri Evangelista[2], Riccardo Ferrazzoli[2], Shuichi Gunji[45], Kiyoshi Hayashida[46]†, Jeremy Heyl[47], Wataru Iwakiri[48], Svetlana G. Jorstad[49,50], Vladimir Karas[3], Jeffery J. Kolodziejczak[29], Fabio La Monaca[2], Ioannis Liodakis[51], Simone Maldera[25], Alberto Manfreda[38], Alan P. Marscher[49], Herman L. Marshall[52], Ikuyuki Mitsuishi[53], Chi-Yung Ng[54], Stephen L. O'Dell[29], Chiara Oppedisano[25], Alessandro Papitto[35], George G. Pavlov[55], Abel L. Peirson[22], Matteo Perri[36,35], Melissa Pesce-Rollins[38], Maura Pilia[37], Andrea Possenti[37], Simonetta Puccetti[36], Brian D. Ramsey[29], Roger W. Romani[22], Carmelo Sgrò[38], Patrick Slane[56], Gloria Spandre[38], Toru Tamagawa[15], Fabrizio Tavecchio[57], Roberto Taverna[58], Yuzuru Tawara[53], Nicholas E. Thomas[29], Alessio Trois[37], Sergey Tsygankov[4,6], Roberto Turolla[58,59], Jacco Vink[60], Kinwah Wu[59], Fei Xie[2,61], Silvia Zane[59]

A black hole x-ray binary (XRB) system forms when gas is stripped from a normal star and accretes onto a black hole, which heats the gas sufficiently to emit x-rays. We report a polarimetric observation of the XRB Cygnus X-1 using the Imaging X-ray Polarimetry Explorer. The electric field position angle aligns with the outflowing jet, indicating that the jet is launched from the inner x-ray–emitting region. The polarization degree is 4.01 ± 0.20% at 2 to 8 kiloelectronvolts, implying that the accretion disk is viewed closer to edge-on than the binary orbit. These observations reveal that hot x-ray–emitting plasma is spatially extended in a plane perpendicular to, not parallel to, the jet axis.

Cygnus X-1 (Cyg X-1, also cataloged as HD 226868) is a bright and persistent x-ray source. It is a binary system containing a 21.2 ± 2.2 solar-mass black hole in a 5.6-day orbit with a $40.6^{+7.7}_{-7.1}$ solar-mass star and is located at a distance of $2.22^{+0.18}_{-0.17}$ kiloparsecs (kpc) (*1*). Gas is stripped from the companion star; as it falls in the strong gravitational field of the black hole, it forms an accretion disk that is heated to millions of kelvin. The hot incandescent gas emits x-rays. Previous analyses of the thermal x-ray flux, its energy spectrum, and the shape of the x-ray emission lines have indicated that the black hole in Cyg X-1 spins rapidly, with a dimensionless spin parameter $a > 0.92$ (close to the maximum possible value of 1) (*2*). Cyg X-1 also produces two pencil-shaped outflows of magnetized plasma, called jets, that have been imaged in the radio band (*3*). It is therefore classified as a microquasar, being analogous to much larger radio-loud quasars (supermassive black holes with jets).

Black hole x-ray binaries are observed in states of x-ray emission thought to correspond to different configurations of the accreting matter (*4*). In the soft state, the x-rays are dominated by thermal emission from the accretion disk. The thermal emission is expected to be polarized because x-rays scatter off electrons in the accretion disk (*5*–*7*). In the hard state, the x-ray emission is produced by (single or multiple) scattering of photons (emitted by the accretion disk or electrons in the magnetic field) off electrons in hot coronal gas. Observations constrain the corona to be much hotter ($k_B T_e \sim 100$ keV, where k_B is the Boltzmann constant and T_e is the electron temperature) than the accretion disk ($k_B T_d \sim 0.1$ keV, where T_d is the disk temperature). The shape of the corona and its location with respect to the accretion disk are both debated (*4*, *8*) but could be constrained by x-ray polarimetry (*9*). Reflection of x-rays emitted by the corona off the accretion disk produces an emission component that includes the iron Kα fluorescence line at ~6.4 keV, which can constrain the velocity of the accretion disk gas orbiting the black hole and the time dilation close to the black hole. This reflection component is also expected to be polarized (*10*, *11*).

We performed x-ray polarimetric observations of Cyg X-1 using the Imaging X-ray Polarimetry Explorer (IXPE) space telescope (*12*). Theoretical predictions of the Cyg X-1 polarization degree (in the 2–8 keV IXPE band) were ~1% or lower, depending on the emission state (*6*, *7*, *9*, *13*). These predictions used an inclination angle (the angle between the black hole spin axis and the line of sight) of $i = 27.^{\circ}5 \pm 0.^{\circ}8$ inferred from optical observations of the binary system (*1*). Earlier polarization observations with the Eighth Orbiting Solar Observatory (OSO-8) space telescope gave a polarization degree of 2.44 ± 1.07% and a polarization angle (measured on the plane of the sky from north to east) of −18° ± 13° at 2.6 keV (*14*, *15*) and a nondetection at higher energies (*16*). IXPE observed Cyg X-1 from 15 to 21 May 2022 with an exposure time of ~242 kiloseconds (ks). The IXPE 2–8 keV observations were coordinated with simultaneous x-ray and gamma-ray observations by other space telescopes covering the energy range 0.2–250 keV, including the Neutron Star Interior Composition Explorer (NICER, 0.2–12 keV), the Nuclear Spectroscopic Telescope Array (NuSTAR, 3–79 keV), the Swift X-ray Telescope (XRT, 0.2–10 keV), the Astronomical Roentgen Telescope–X-ray Concentrator (ART-XC, 4–30 keV) of the Spectrum-Röntgen-Gamma observatory (SRG), and the INTEGRAL Soft Gamma-Ray Imager (ISGRI, 30–80 keV) on the International Gamma-Ray Astrophysics Laboratory (INTEGRAL) (*17*). Simultaneous optical observations were performed with the Double Image Polarimeter 2 (DIPol-2) instrument mounted on the Tohoku 60-cm telescope at the Haleakala Observatory, Hawaii, and the Robotic Polarimeter (RoboPol) at the 1.3-m telescope of the Skinakas Observatory, Greece (*17*).

During the observation campaign, Cyg X-1 was highly variable over the entire 0.2–250 keV energy range (fig. S1). The source was in the hard x-ray state with a photon index of 1.6 (table S5) and a 0.2–250 keV luminosity of 1.1% of the Eddington luminosity (the luminosity at which the radiation pressure on electrons equals the gravitational pull on the ions of the accreted material). We detected linear polarization in the IXPE data with >20σ statistical confidence (where σ is the standard deviation) (Fig. 1 and fig. S3), measuring a 2–8 keV polarization degree of 4.01 ± 0.20% at an electric field position angle of $-20.^{\circ}7 \pm 1.^{\circ}4$. The polarization degree and angle are consistent with the previous results of OSO-8 at 2.6 keV (*15*). Evidence for an increase in the polarization degree with energy

(Fig. 1 and fig. S5) is significant at the 3.4σ level (*17*). We find a 2.4σ indication that the polarization degree increases with the source flux (fig. S6).

We find no evidence that the polarization depends on the orbital phase of the binary system (fig. S7). This excludes the possibility that the observed x-ray polarization originates from the scattering of x-ray photons off the companion star or its wind and shows that these effects do not measurably affect the polarization properties.

We calculated a suite of emission models and compared them with the observations (*17*). We estimate that >90% of the x-rays come from the inner ~2000-km-diameter region surrounding the ~60-km-diameter black hole. The x-ray polarization angle aligns with the billion-kilometer-scale radio jet to within ~5° (Fig. 2).

We decomposed the broadband energy spectra observed simultaneously with IXPE, NICER, NuSTAR, and INTEGRAL into a multitemperature black-body component (thermal emission from the accretion disk), a power-law component (from multiple Compton scattering events in the corona), emission reflected off the accretion disk, and emission from more distant stationary plasma (fig. S8) (*17*). We find that the coronal emission strongly dominates in the IXPE energy band, contributing ~90% of the observed flux. The accretion disk and reflected emission components contribute <1% and ~10% of the emission, respectively. Therefore, our polarization measurements are likely to be dominated by the coronal emission.

We analyzed the optical data at multiple wavelengths (*17*), finding an intrinsic optical polarization degree of ~1% and polarization angle of −24°. The uncertainties on these results are dominated by systematic effects related to the choice of polarization reference stars and are ±0.1% on the polarization degree and ±13° on the polarization direction (figs. S11 to S13 and table S4). The optical polarization direction is thought to indicate the orientation of the orbital axis projected onto the sky (*18*). We find that it aligns with the x-ray polarization direction and the radio jet.

The alignment of the x-ray polarization with the radio jet indicates that the inner x-ray–emitting region is directly related to the radio jet. If the x-ray polarization is perpendicular to the inner accretion disk plane, as favored in our models (*17*), this implies that the inner accretion disk is perpendicular to the radio jet, at least on the plane of the sky. This is consistent with the hypothesis that jets of microquasars (and, by extension, of quasars) are launched perpendicular to the inner accretion flow (*19*).

Figure 3 compares our observed polarization with theoretical predictions made using models of the corona (*17*). We find that the only models that are consistent with the observations are those in which the coronal plasma is extended perpendicular to the jet axis, and therefore probably parallel to the accretion disk. In these models, repeated scatterings in the plane of the corona polarize the x-rays perpendicular to that plane. Two models are consistent with our observations: (i) a hot corona sandwiching the accretion disk (*20*), as predicted by numerical accretion disk simulations (*21*); or (ii) a composite accretion flow with a truncated cold disk that is geometrically thin and optically thick and an inner laterally extended region (geometrically thick but optically thin) of hot plasma, possibly produced by evaporation of the cold disk (*22*). If the jet is launched from the inner, magnetized region of the disk, the jet carrying away disk angular momentum could leave behind a radially extended hot and optically thin corona (*23*).

The polarization data rule out models in which the corona is a narrow plasma column or cone along the jet axis, or consists of two compact regions above and below the black hole. Our modeling of these scenarios accounts for the effect of the coronal emission reflecting off the accretion disk (*17*). These models predict polarization degree well below the observed values. Models that produce high polarization degree predict polarization directions close to perpendicular to the jet axis, a decreasing polarization degree with energy, or both, and therefore disagree with the observations.

In our favored corona models, the high polarization degree we observe requires that the x-ray bright region is seen at a higher inclination than the ~27° inclination of the binary orbit. Sandwich corona models involving the

[1]Department of Physics and McDonnell Center for the Space Sciences, Washington University in St. Louis, St. Louis, MO 63130, USA. [2]Istituto di Astrofisica e Planetologia Spaziali, Istituto Nazionale di Astrofisica (INAF), 00133 Roma, Italy. [3]Astronomical Institute of the Czech Academy of Sciences, 14100 Praha 4, Czech Republic. [4]Department of Physics and Astronomy, 20014 University of Turku, Turku, Finland. [5]Nordic Institute for Theoretical Physics (Nordita), Kungliga Tekniska Högskolan (KTH) Royal Institute of Technology and Stockholm University, SE-106 91 Stockholm, Sweden. [6]Space Research Institute of the Russian Academy of Sciences, Moscow 117997, Russia. [7]School of Mathematics, Statistics, and Physics, Newcastle University, Newcastle upon Tyne NE1 7RU, UK. [8]Dipartimento di Matematica e Fisica, Università degli Studi Roma Tre, 00146 Roma, Italy. [9]Division of Physics, Mathematics and Astronomy, California Institute of Technology, Pasadena, CA 91125, USA. [10]Center for Space Sciences and Technology, University of Maryland, Baltimore County, Baltimore, MD 21250, USA. [11]NASA Goddard Space Flight Center (GSFC), Greenbelt, MD 20771, USA. [12]Center for Research and Exploration in Space Science and Technology, NASA-GSFC, Greenbelt, MD 20771, USA. [13]Rikagaku Kenkyūjyo (RIKEN) Cluster for Pioneering Research, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan. [14]Centre national de la recherche scientifique, Observatoire Astronomique de Strasbourg, Université de Strasbourg, Unité Mixte de Recherche 7550, 67000 Strasbourg, France. [15]Astronomical Institute, Charles University, 18000 Prague, Czech Republic. [16]National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China. [17]Leibniz-Institut für Sonnenphysik, 79104 Freiburg, Germany. [18]Istituto Ricerche Solari (IRSOL) Aldo e Cele Daccò, Faculty of Informatics, Università della Svizzera italiana, 6605 Locarno, Switzerland. [19]Euler Institute, Faculty of Informatics, Università della Svizzera italiana, 6962 Lugano, Switzerland. [20]Institute of Astrophysics, Foundation for Research and Technology–Hellas, 71110 Heraklion, Greece. [21]Department of Physics, University of Crete, 70013 Heraklion, Greece. [22]Department of Physics and Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, Stanford, CA 94305, USA. [23]Department of Astronomy, University of Michigan, Ann Arbor, MI 48109, USA. [24]School of Sciences, Tohoku University, Aoba-ku, 980-8578 Sendai, Japan. [25]Istituto Nazionale di Fisica Nucleare, Sezione di Torino, 10125 Torino, Italy. [26]Agenzia Spaziale Italiana (ASI), 00133 Roma, Italy. [27]Hiroshima Astrophysical Science Center, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8526, Japan. [28]Institut de Planétologie et d'Astrophysique de Grenoble (IPAG), Université Grenoble Alpes, Centre national de la recherche scientifique, 38000 Grenoble, France. [29]NASA Marshall Space Flight Center, Huntsville, AL 35812, USA. [30]Dr. Karl Remeis Observatory, Erlangen Centre for Astroparticle Physics, Universität Erlangen-Nürnberg, 96049 Bamberg, Germany. [31]Dipartimento di Fisica, Università degli Studi di Roma "Tor Vergata," 00133 Roma, Italy. [32]Istituto Nazionale di Fisica Nucleare, Sezione di Roma "Tor Vergata," 00133 Roma, Italy. [33]Department of Astronomy, University of Maryland, College Park, MD 20742, USA. [34]Instituto de Astrofísica de Andalucía, 18008 Granada, Spain. [35]INAF Osservatorio Astronomico di Roma, 00078 Monte Porzio Catone, Roma, Italy. [36]Space Science Data Center, ASI, 00133 Roma, Italy. [37]INAF Osservatorio Astronomico di Cagliari, 09047 Selargius, Cagliari, Italy. [38]Istituto Nazionale di Fisica Nucleare, Sezione di Pisa, 56127 Pisa, Italy. [39]Dipartimento di Fisica, Università di Pisa, 56127 Pisa, Italy. [40]Dipartimento di Fisica, Università degli Studi di Torino, 10125 Torino, Italy. [41]INAF Osservatorio Astrofisico di Arcetri, 50125 Firenze, Italy. [42]Dipartimento di Fisica e Astronomia, Università degli Studi di Firenze, 50019 Sesto Fiorentino, Firenze, Italy. [43]Istituto Nazionale di Fisica Nucleare, Sezione di Firenze, 50019 Sesto Fiorentino, Firenze, Italy. [44]Institut für Astronomie und Astrophysik, Universität Tübingen, 72076 Tübingen, Germany. [45]Department of Physics, Yamagata University, 1-4-12 Kojirakawa-machi, Yamagata-shi 990-8560, Japan. [46]Department of Earth and Space Science, Osaka University, 1-1 Yamadaoka, Suita, Osaka 565-0871, Japan. [47]Department of Physics and Astronomy, University of British Columbia, Vancouver, BC V6T 1Z4, Canada. [48]Department of Physics, Faculty of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan. [49]Institute for Astrophysical Research, Boston University, Boston, MA 02215, USA. [50]Department of Astrophysics, St. Petersburg State University, Petrodvoretz, 198504 St. Petersburg, Russia. [51]Finnish Centre for Astronomy with the European Southern Observatory (ESO), 20014 University of Turku, Turku, Finland. [52]Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [53]Division of Particle and Astrophysical Science, Graduate School of Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8602, Japan. [54]Department of Physics, The University of Hong Kong, Pokfulam, Hong Kong. [55]Department of Astronomy and Astrophysics, Pennsylvania State University, University Park, PA 16802, USA. [56]Center for Astrophysics, Harvard & Smithsonian, Cambridge, MA 02138, USA. [57]INAF Osservatorio Astronomico di Brera, 23807 Merate, Lecco, Italy. [58]Dipartimento di Fisica e Astronomia, Università degli Studi di Padova, 35131 Padova, Italy. [59]Mullard Space Science Laboratory, University College London, Holmbury St Mary, Dorking, Surrey RH5 6NT, UK. [60]Anton Pannekoek Institute for Astronomy, University of Amsterdam, 1098 XH Amsterdam, Netherlands. [61]Guangxi Key Laboratory for Relativistic Astrophysics, School of Physical Science and Technology, Guangxi University, Nanning 530004, China.

***Corresponding author. Email: krawcz@wustl.edu (H.K.); fabio.muleri@inaf.it (F.Mu.); dovciak@astro.cas.cz (M.D.); alexandra.veledina@utu.fi (A.V.)**

†Deceased.

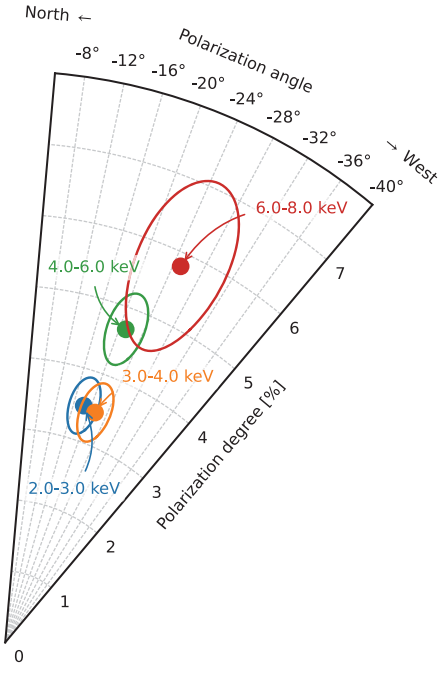


Fig. 1. Energy-dependent x-ray polarization of Cyg X-1. The polarization degree and polarization angle, derived from the IXPE observations, are shown for four energy bands (labeled and in different colors). The ellipses denote the 68.3% confidence regions.

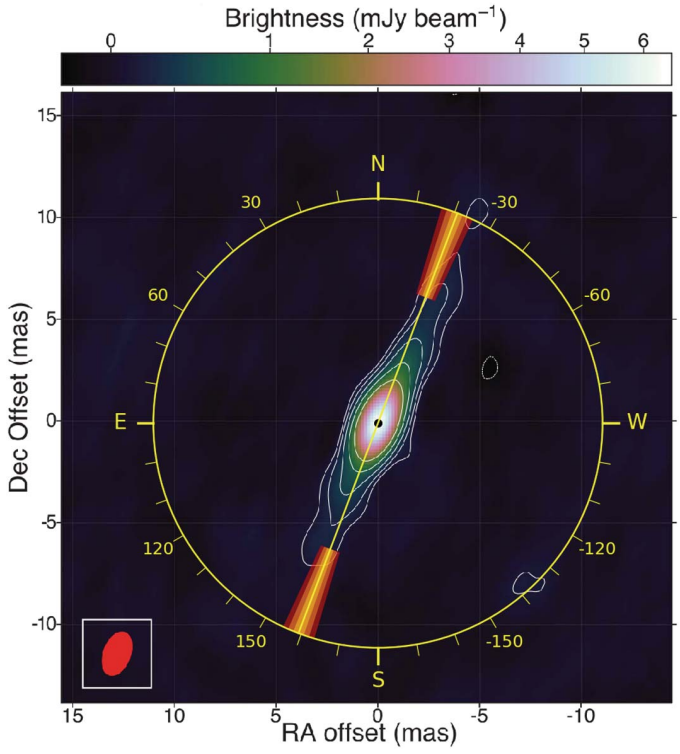


Fig. 2. Comparison of the x-ray polarization direction with the radio jet. The 2–8 keV electric vector position angle is shown with the yellow line, and the 1σ, 2σ, and 3σ confidence regions are given by the orange-to-red shading. The background image is a radio observation of the jet (*1*). We infer (see text) that most x-rays are emitted by a ~2000-km-diameter region surrounding the ~60-km-diameter black hole, far smaller than the resolution of the radio image (which is indicated by the red ellipse). The coordinate offsets in right ascension (RA) and declination (Dec) (J2000 equinox) are in units of milliarcseconds (mas). The color scale shows the radio flux in milli-Jansky, with 1 Jansky being 10^{-26} W m^{-2} Hz^{-1}.

Compton scattering of disk photons with initial energies of ~0.1 keV require inclinations exceeding 65°. Truncated disk models invoking Compton scattering of the disk or internally generated lower-energy (~1–10 eV) synchrotron photons (*24*) can reproduce the observed polarization degree for inclinations of >45°. In comparison to the models with disk photons, the larger number of scatterings required to energize lower-energy synchrotron photons to kiloelectronvolt energies results in higher polarization degree in the IXPE energy band (fig. S9) (*17*).

Although the x-ray polarization, optical polarization, and radio jet approximately align in the plane of the sky, the inclination of the x-ray bright region exceeds that of the binary orbit, implying that the inner accretion flow is seen more edge-on than the binary orbit. Because the bodies of a stellar system typically orbit and spin around the same axis (as do most planets in the Solar System), we consider potential explanations for the mismatch between the inner accretion disk inclination and the orbital inclination.

Stellar-mass black holes are formed during supernovae. The supernova that occurred in Cyg X-1 might have left the black hole with a misaligned spin. Gravitational effects could align the inner accretion flow angular momentum vector with the black hole spin vector (*25*). In this scenario, aligning the inner accretion disk angular momentum vector with the black hole spin vector would also align the radio jet produced by the inner accretion disk with the black hole spin vector. Several, but not all, analyses of Cyg X-1 reflected emission spectra give inclinations consistent with our $i > 45°$ constraint (*26*, *27*).

An alternative explanation for the large inclination of the x-ray–emitting region invokes the precession of the inner accretion flow with a period much longer than the orbital period (*28*). From our analysis of a 2–4 keV long-term x-ray light curve, we infer that the IXPE observations were performed close to the maximum inner disk inclination (fig. S2) (*17*). We tested the hypothesis that the inner flow precesses with an amplitude of $\gtrsim 17.^{\circ}5$ by performing an additional 86-ks IXPE target of opportunity observation of Cyg X-1 from 18 to 20 June 2022, 33 days after the May observations, which corresponds to half of the current superorbital period (*17*). If this hypothesis is correct, we expect the polarization degree to drop from 4.01 ± 0.20% to ≪1% owing to the inclination changing from $i > 45°$ in May to $i \lesssim 10°$ in June. The observations showed the source in the same hard state with a 2–8 keV polarization degree and angle of 3.84 ± 0.31% and $-25.^{\circ}7 \pm 2.^{\circ}3$, respectively (fig. S4) (*17*). The polarization degree remained constant (within the statistical uncertainties) between the May and June observations. We therefore disfavor the hypothesis that precession of the inner accretion flow leads to the high polarization degree of the May observation. The combined May and June polarization degree and angle are 3.95 ± 0.17% and $-22.^{\circ}2 \pm 1.^{\circ}2$, respectively (fig. S4) (*17*).

In previous work, others have argued that optically thin synchrotron emission from the base of the jet could contribute up to 5% to the Cyg X-1 x-ray emission in the hard state (*29*, *30*). Synchrotron emission from electrons gyrating around magnetic field lines is polarized perpendicular to those field lines. Our observation of the x-rays being polarized parallel to the jet axis would require synchrotron emission from a toroidal magnetic field, wound around the jet axis. For this magnetic field geometry, seen at an inclination of $27.^{\circ}5$, the theoretical upper limit on the polarization degree of the synchrotron emission is 8% (*31*). The jet thus contributes <0.4% of the observed polarization degree. If the almost-constant jet

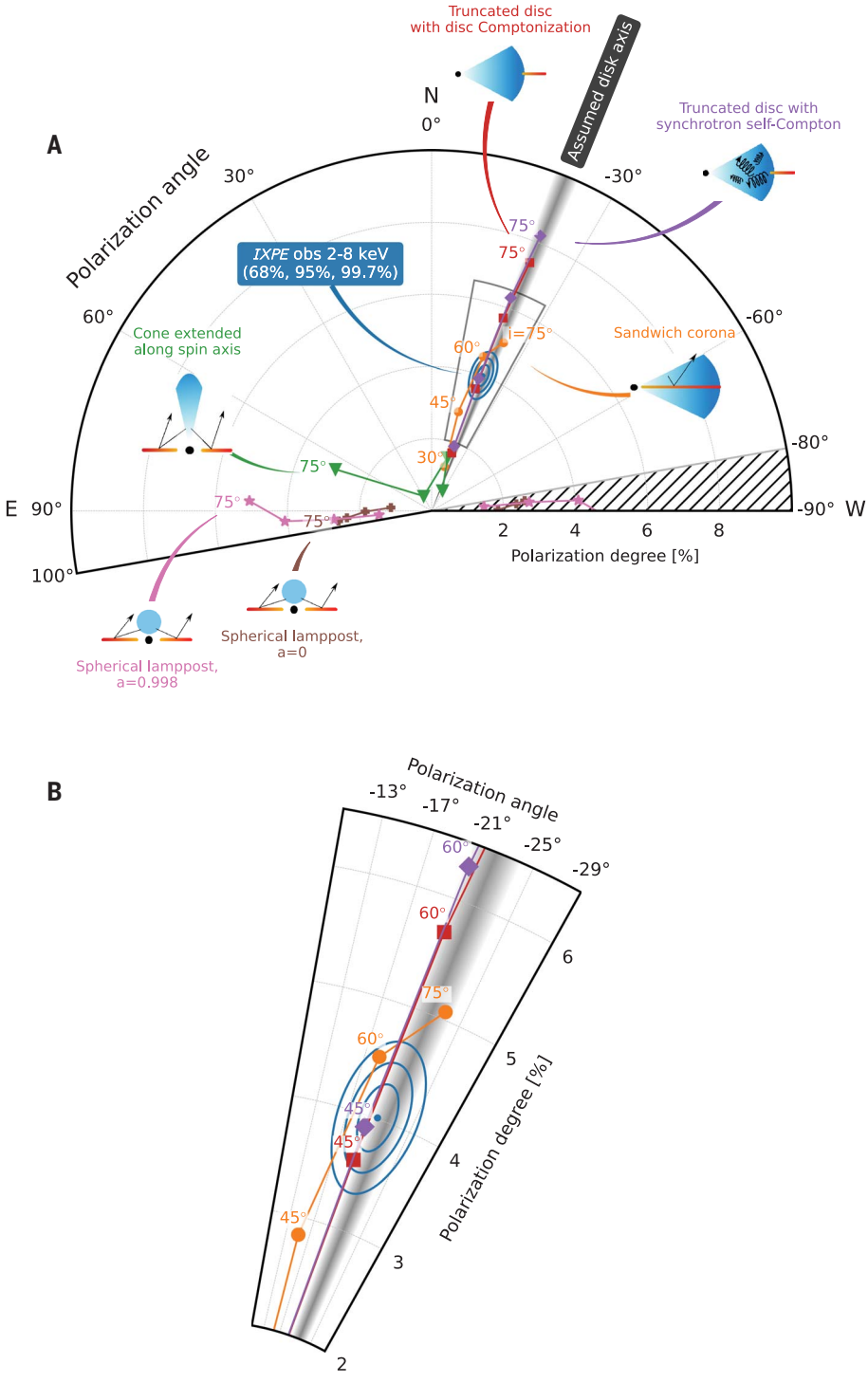


Fig. 3. Comparison of the observed 2–8 keV polarization degree and angle with model predictions. (**A**) The blue dot shows the polarization degree and angle, with the blue ellipses indicating the 68, 95, and 99.7% confidence levels (equivalent to 1σ, 2σ, and 3σ, respectively). Model predictions assume that the inner disk spin axis has position angle of −22° (consistent with the radio jet), and that the inner disk angular momentum vector points away from the observer (as does the orbital angular momentum vector) (*1*). The gray band shows the uncertainty of the radio jet orientation; we adopt this as the uncertainty of the disk spin axis in all models. Each colored line shows the model results for each chosen corona geometry, with symbols indicating different values as a function of the inner disk inclination *i*. Inset diagrams schematically depict the assumed black hole (black), corona (blue), and accretion disk (orange-red) configurations. Black arrows indicate photon paths. Models with coronae extending parallel to the inner accretion disk can match the IXPE observations, but coronae located or extending along the spin axis of the inner accretion disk cannot. The position angles are shown from −80° to +100° (instead of −90° to +90°) to clarify the models that straddle the ±90° borders. (**B**) A zoom into the region around the measured value, marked with the gray box in (A).

emission was the main source of the observed polarization, we would expect that a rise in the x-ray flux from the inner accretion flow would lead to an overall smaller polarization degree—contrary to the observed trend (fig. S6).

The polarized x-rays from the immediate surroundings of the black hole carry the imprint of the geometry of the emitting gas. We conclude that the x-ray bright plasma is extended perpendicular to the radio jet. The high observed polarization degree either implies a more edge-on viewing geometry than given by the optical data, or it suggests that unidentified physical effects are responsible for production of the x-rays in accreting black hole systems.

REFERENCES AND NOTES

1. J. C. A. Miller-Jones *et al.*, *Science* **371**, 1046–1049 (2021).
2. L. Gou *et al.*, *Astrophys. J.* **790**, 29 (2014).
3. A. M. Stirling *et al.*, *Mon. Not. R. Astron. Soc.* **327**, 1273–1278 (2001).
4. C. Done, M. Gierliński, A. Kubota, *Astron. Astrophys. Rev.* **15**, 1–66 (2007).
5. P. A. Connors, T. Piran, R. F. Stark, *Astrophys. J.* **235**, 224 (1980).
6. L.-X. Li, R. Narayan, J. E. McClintock, *Astrophys. J.* **691**, 847–865 (2009).
7. J. D. Schnittman, J. H. Krolik, *Astrophys. J.* **701**, 1175–1187 (2009).
8. C. Bambi *et al.*, *Space Sci. Rev.* **217**, 65 (2021).
9. J. D. Schnittman, J. H. Krolik, *Astrophys. J.* **712**, 908–924 (2010).

10. G. Matt, *Mon. Not. R. Astron. Soc.* **260**, 663–674 (1993).
11. J. Poutanen, K. N. Nagendra, R. Svensson, *Mon. Not. R. Astron. Soc.* **283**, 892–904 (1996).
12. M. C. Weisskopf *et al.*, *J. Astron. Telesc. Instrum. Syst.* **8**, 026002 (2022).
13. H. Krawczynski, B. Beheshtipour, *Astrophys. J.* **934**, 4 (2022).
14. M. C. Weisskopf *et al.*, *Astrophys. J.* **215**, L65 (1977).
15. K. S. Long, G. A. Chanan, R. Novick, *Astrophys. J.* **238**, 710 (1980).
16. M. Chauvin *et al.*, *Nat. Astron.* **2**, 652–655 (2018).
17. Materials and methods are available as supplementary materials.
18. J. C. Kemp, M. S. Barbour, T. E. Parker, L. C. Herman, *Astrophys. J.* **228**, L23–L27 (1979).
19. M. C. Begelman, R. D. Blandford, M. J. Rees, *Rev. Mod. Phys.* **56**, 255–351 (1984).
20. F. Haardt, L. Maraschi, *Astrophys. J.* **380**, L51 (1991).
21. B. E. Kinch, J. D. Schnittman, S. C. Noble, T. R. Kallman, J. H. Krolik, *Astrophys. J.* **922**, 270 (2021).
22. F. Meyer, E. Meyer-Hofmeister, *Astron. Astrophys.* **288**, 175–182 (1994).
23. P. O. Petrucci, J. Ferreira, G. Henri, J. Malzac, C. Foellmi, *Astron. Astrophys.* **522**, A38 (2010).
24. A. Veledina, J. Poutanen, I. Vurm, *Mon. Not. R. Astron. Soc.* **430**, 3196–3212 (2013).
25. J. M. Bardeen, J. A. Petterson, *Astrophys. J.* **195**, L65 (1975).
26. J. A. Tomsick *et al.*, *Astrophys. J.* **780**, 78 (2014).
27. M. L. Parker *et al.*, *Astrophys. J.* **808**, 9 (2015).
28. P. Lachowicz, A. A. Zdziarski, A. Schwarzenberg-Czerny, G. G. Pooley, S. Kitamoto, *Mon. Not. R. Astron. Soc.* **368**, 1025–1039 (2006).
29. D. M. Russell, T. Shahbaz, *Mon. Not. R. Astron. Soc.* **438**, 2083–2096 (2014).
30. A. A. Zdziarski, P. Pjanka, M. Sikora, Ł. Stawarz, *Mon. Not. R. Astron. Soc.* **442**, 3243–3255 (2014).
31. M. Lyutikov, V. I. Pariev, D. C. Gabuzda, *Mon. Not. R. Astron. Soc.* **360**, 869–891 (2005).
32. P. Thalhammer, J. Wilms, N. Rodriguez Cavero, X-ray observations of black hole binary Cyg X-1 with INTEGRAL, version 1, Zenodo (2022); https://doi.org/10.5281/zenodo.7140274.
33. V. Kravtsov *et al.*, Optical polarimetric observations of black hole binary Cyg X-1 with DIPol-2, version 1, Zenodo (2022); https://doi.org/10.5281/zenodo.7108247.
34. D. Blinov, S. Kiehlmann, N. Mandarakas, R. Skalidis, Optical polarimetric observations of the black hole binary star Cyg X-1 with RoboPol, version 1, Zenodo (2022); https://doi.org/10.5281/zenodo.7127802.
35. W. Zhang, M. Dovčiak, M. Bursa, *Astrophys. J.* **875**, 148 (2019).
36. A. Veledina, J. Poutanen, Polarization of Comptonized emission in slab geometry, version 1, Zenodo (2022); https://doi.org/10.5281/zenodo.7116125.

ACKNOWLEDGMENTS

We thank J. Miller-Jones, J. Orosz, and A. Zdziarski for very helpful discussions of the optical constraints on the orbital inclination of Cyg X-1 and optical position angles. We are grateful to three anonymous referees, whose excellent comments contributed to strengthening the paper. We thank T. Maccarone for emphasizing that stellar wind absorption may modify the jet orientation measurement results. This work is based on observations made with the IXPE mission, a joint US and Italian mission. The US contribution to the IXPE mission is supported by NASA and led and managed by its Marshall Space Flight Center, with industry partner Ball Aerospace (contract NNM15AA18C). The Italian contribution to the IXPE mission is supported by the Italian Space Agency (ASI) through contract ASI-OHBI-2017-12-I.0, agreements ASI-INAF-2017-12-H0 and ASI-INFN-2017.13-H0, and its Space Science Data Center (SSDC) with agreements ASI-INAF-2022-14-HH.0 and ASI-INFN 2021-43-HH.0; and by INAF and the Istituto Nazionale di Fisica Nucleare (INFN) in Italy. This research used data and software products or online services provided by the IXPE Team (Marshall Space Flight Center, the SSDC of the Italian Space Agency, the INAF, and INFN), as well as the High-Energy Astrophysics Science Archive Research Center (HEASARC), at NASA Goddard Space Flight Center. We thank the NICER, NuSTAR, INTEGRAL, Swift, and SRG/ART-XC teams and Science Operation Centers for their support of this observation campaign. DIPol-2 is a joint effort between University of Turku (Finland) and Leibniz Institut für Sonnenphysik (Germany). We are grateful to the Institute for Astronomy, University of Hawaii, for allocating observing time for the DIPol-2 polarimeter, and to the Skinakas Observatory for performing the observations with the RoboPol polarimeter at their 1.3-m telescope. **Funding:** H.K. acknowledges NASA support under grants 80NSSC18K0264, 80NSSC22K1291, 80NSSC21K1817, and NNX16AC42G. F.Mu., J.R., S.B., S.F., A.R., P.So., E.D.M., E.Co., A.D.M., G.M., Y.E., R.F., F.L.M., M.Pe., and A.T. were funded through contract ASI-INAF-2017-12-H0. L.B., R.Bo., R.Be., A.Br., L.L., S.Ca., S.M., A.Man., C.O., M.P.-R., C.S., and G.S. were funded by the ASI through contracts ASI-INFN-2017.13-H0 and ASI-INFN 2021-43-HH.0. M.Pi. was funded through contract ASI-INAF-2022-14-HH.0. I.A. acknowledges support from MICINN (Ministerio de Ciencia e Innovación) Severo Ochoa award for the IAA-CSIC (SEV-2017-0709) and through grants AYA2016-80889-P and PID2019-107847RB-C44. M.D., J.S., and V.Ka. acknowledge support from GACR (Grantová agentura České republiky) project 21-06825X and institutional support from the Astronomical Institute of the Czech Academy of Sciences (RVO:67985815). J.A.G. acknowledges support from NASA grant 80NSSC20K0540. J.Pod. acknowledges support from Charles University project GA UK No. 174121 and from the Barrande Fellowship Programme of the Czech and French governments. A.V., J.Pou., and S.S.T. acknowledge support from Russian Science Foundation grant 20-12-00364 and the Academy of Finland grants 333112, 347003, 349144, and 349906. M.N. acknowledges support from NASA under award number 80GSFC21M0002. T.K. is supported by JSPS KAKENHI Grant Number JP19K03902. P.-O.P. acknowledges support from the High Energy National Programme (PNHE) of Centre national de la recherche scientifique (CNRS) and from the French space agency (CNES) as well as from the Barrande Fellowship Programme of the Czech and French governments. D.B., S.K., N.M., and R.S. acknowledge support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program under grant agreement no. 771282. V.Kr. thanks Vilho, Yrjö and Kalle Väisälä Foundation. P.T. and J.W. acknowledge funding from Bundesministerium für Wirtschaft and Klimaschutz under Deutsches Zentrum für Luft- und Raumfahrt grant 50 OR 1909. A.I. acknowledges support from the Royal Society. J.H. acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), funding reference number 5007110, and the Canadian Space Agency. S.G.J. and A.P.M. are supported in part by National Science Foundation grant AST-2108622, by NASA Fermi Guest Investigator grant 80NSSC21K1917, and by NASA Swift Guest Investigator grant 80NSSC22K0537. C.-Y.N. is supported by a General Research Fund of the Hong Kong Government under grant number HKU 17305419. P.Sl. acknowledges support from NASA Contract NAS8-03060. **Author contributions:** H.K., F.Mu., M.D., A.V., N.R.C., J.S., A.I., G.M., J.A.G., V.L., and J.Pou. participated in the planning of the observation campaign and the analysis and modeling of the data. M.N., T.K., J.Pod., J.R., and W.Z. contributed to the analysis or modeling of the data. A.V.B., V.Kr., S.V.B., M.K., T.S., D.B., S.K., N.M., and R.S. contributed to the optical polarimetric data. J.M.M. and P.D. contributed the Swift results; J.W. and P.T. the INTEGRAL results; and A.A.L., S.V.M., and A.N.S. the SRG/ART-XC results. S.B., F.C., N.D.L., L.L., A.Mar., T.M., N.O., A.R., P.-O.P, P.So., A.F.T., F.To., M.C.W., and S.Zh. contributed to the discussion of the results. F.Ma. and S.F. served as internal referees. All other authors contributed to the design and science case of the IXPE mission and to planning the observations used in this paper. All authors provided input and comments on the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The May and June IXPE observations are available at https://heasarc.gsfc.nasa.gov/FTP/ixpe/data/obs/01/01002901/ and https://heasarc.gsfc.nasa.gov/FTP/ixpe/data/obs/01/01250101/, respectively. The NICER data are available at https://heasarc.gsfc.nasa.gov/docs/nicer/nicer_archive.html under ObsIDs 5100320101, 5100320102, 5100320103, 5100320104, 5100320105, 5100320106, and 5100320107. The NuSTAR data are available at https://heasarc.gsfc.nasa.gov/db-perl/W3Browse/w3table.pl?tablehead=name%3Dnumaster&Action=More+Options under ObsIDs 30702017002, 30702017004, and 30702017006. The SWIFT XRT data are available at https://heasarc.gsfc.nasa.gov/cgi-bin/W3Browse/swift.pl under ObsIDs 00034310009, 00034310010, 00034310011, 00034310012, 00034310013, and 00034310014. The extracted INTEGRAL ISGRI data are archived at Zenodo (*32*). The SRG ART-XC data are available at ftp://hea.iki.rssi.ru/public/SRG/ART-XC/data/Cygnus_X-1/. The MAXI light curves are available at http://maxi.riken.jp/star_data/J1958+352/J1958+352.html. The raw DIPol-2 and RoboPol data are archived at Zenodo (*33*, *34*). The KERRC code (*13*) is available at https://gitlab.com/krawcz/kerrc-x-ray-fitting-code.git. The MONK code (*35*) is available at https://projects.asu.cas.cz/zhang/monk. The ixpeobssim software is available at https://github.com/lucabaldini/ixpeobssim and documented at https://ixpeobssim.readthedocs.io. Our derived x-ray polarization measurements are listed in tables S1 and S2, and the optical polarization measurements are listed in table S4. The numerical results of our model fitting are listed in table S5. Our models of polarized emission in the truncated disk geometry are archived at Zenodo (*36*). **License information:** Copyright © 2022 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.add5399
Materials and Methods
Figs. S1 to S12
Tables S1 to S5
References (*37*–*79*)

Submitted 18 June 2022; accepted 17 October 2022
Published online 3 November 2022
10.1126/science.add5399

RAINFALL EXTREMES

Intensification of subhourly heavy rainfall

Hooman Ayat[1,2]*, Jason P. Evans[1,2], Steven C. Sherwood[1,2], Joshua Soderholm[3]

Short-duration rainfall extremes can cause flash flooding with associated impacts. Previous studies of climate impacts on extreme precipitation have focused mainly on daily rain totals. Subdaily extremes are often generated in small areas that can be missed by gauge networks or satellites and are not resolved by climate models. Here, we show a robust positive trend of at least 20% per decade in subhourly extreme rainfall near Sydney, Australia, over 20 years, despite no evidence of trends at hourly or daily scales. This trend is seen consistently in storms tracked using multiple independent ground radars, is consistent with rain-gauge data, and does not appear to be associated with known natural variations. This finding suggests that subhourly rainfall extremes may be increasing substantially faster than those on more widely reported time scales.

Climate change is expected to change the intensity and frequency of heavy rainfall across the world (*1*–*3*) by increasing humidity and changing the large-scale atmospheric circulation and convective dynamics of storms (*4*). Despite advances in understanding the effects of a warming climate on rainfall extremes at daily (or longer) scales, according to the sixth report of the Intergovernmental Panel on Climate Change (IPCC), there is very low confidence about changes in short-duration (subdaily) extreme precipitation, which is often responsible for destructive natural hazards like flash floods, landslides, and debris flows in both urban and rural areas (*5*–*7*). Improved knowledge of changes in heavy, short-duration rainfall is vital for effective climate adaptation (*6*–*8*) and to reduce the vulnerability of cities (*6*).

Current studies have shown upward trends in extreme rainfall at daily scales in many places around the globe and suggest typical increases similar to the Clausius-Clapeyron rate of ~7% K^{-1} (*1*, *8*–*11*). Super–Clausius-Clapeyron scaling (>7% K^{-1}) is reported for subdaily rainfall extremes in some regions like Australia (*1*, *12*, *13*), Germany (*14*), Netherlands (*15*–*17*), South Korea (*18*), and Hong Kong (*19*). However, it is uncertain whether these observed scaling rates can be used as a basis for projecting future changes to rainfall extremes because this approach does not explicitly include other contributing factors such as changes in atmospheric stability, storm structural dynamics (*20*), and large-scale circulation (*4*, *17*, *21*).

A growing number of studies have examined trends at subdaily and/or subhourly scales using gauge stations during the past decades in various regions across the globe. Although these studies reported positive trends in subdaily rainfall extremes at some locations in North America (*22*, *23*), Australia (*1*, *2*, *7*, *24*–*30*), Europe (*31*–*33*), Southeast Asia (*34*), and parts of China (*35*), most stations (within a network) showed weak or no trends, probably because of their questionable representativeness in sampling short-duration (subdaily or subhourly) rainfall extremes. These extremes often take place over small areas (e.g., subkilometer) and time (e.g., subhourly) scales that are rarely resolved by gauge networks. The sixth report of the IPCC considered this issue to be one of the main limitations in determining reliable trends in short-duration rainfall extremes (*36*). For example, Lengfeld *et al.* (*36*) showed that only 17.3% of hourly heavy precipitation events in Germany from 2001 to 2018 were captured by the gauge network, whereas 81.8% of daily events were observed.

Regional climate models (RCMs) can predict subdaily rainfall extremes. A few RCM-based studies showed them intensifying over North America (*37*) and Sweden (*38*). However, subdaily rainfall extremes are poorly simulated in RCMs, owing to parameterizing convection rather than explicitly simulating it. Convection-permitting models (CPMs) overcome this limitation and have been run in a few regions (*39*, *40*). However, there are large uncertainties in the fidelity and evaluation of CPMs because of the lack of reliable observations, the spatiotemporal scale mismatch between simulated and observed data, and the reliance of these models on parameterization of microscale processes (*41*, *42*). Similar concerns apply to high-resolution reanalysis datasets (*43*).

The limitations mentioned above suggest that more spatially and temporally complete observational data are required to investigate hourly or shorter-duration high-impact rainfall events. Satellite-based precipitation products can provide precipitation estimation with global coverage at reasonably high spatiotemporal resolution. However, observations in these products have large spatiotemporal gaps that are filled by sophisticated interpolation techniques combined with observations from cloud top temperatures that are converted to precipitation estimates, leading to large biases in heavy precipitation (*44*). Thus, short-duration rainfall extremes are not satisfactorily captured by current satellite products.

The gap between rain-gauge networks, satellite observations, and climate models can be filled by weather-radar data, which provide precipitation measurements at high spatiotemporal resolution (*36*). However, these datasets are subject to uncertainties and inhomogeneity caused by hardware upgrades, calibration, maintenance, attenuation, ground clutter, beam blocking, and/or merging different sensors to cover a larger area, which makes their application to trend analysis challenging (*36*).

Here, we follow a new approach to overcome these limitations and use weather radars to gain higher confidence in changes of short-duration rainfall extremes. We used satellite-calibrated observations from three ground-based radars (Wollongong, Terrey Hills, and Newcastle) covering the greater Sydney, Australia, region (*45*) to study trends in rainfall extremes down to the 10-min scale over the past two decades. Figure 1, A to C, shows the locations of the three radars and their fields of view while observing the same events at a particular time. Using multiple overlapping radars enabled us to limit the data inconsistency issue in the datasets and increase the signal-to-noise ratio.

By using an object-based approach, rain systems in the gridded radar data are considered as moving objects (see the highlighted objects in Fig. 1, A to C). This approach enabled us to locate and observe the peak intensity of storms for every time step, which is not possible using point-based datasets. The object-based method used in this study is called method of object-based diagnostic evaluation (MODE), which has been previously used over Wollongong radar in another study to investigate the climatology of rain systems near Sydney (*46*). Using this technique over two radar- and satellite-based precipitation products over the eastern United States showed that the object-based properties of the storms are not dependent on the observational platform (*44*, *47*). By using this technique, the spatial maximum intensity of the detected objects (hereafter, object maximum intensity) for each time step have been extracted and are paired with their year of occurrence. A 95% quantile regression has been applied over the paired data (for each radar separately) to target the trend of the spatial maximum rain rate of the extreme storms.

We also repeated the quantile regression analysis over the objects captured simultaneously by both the Terrey Hills and Wollongong radars to identify any effect of different radar

[1]Climate Change Research Centre, University of New South Wales, Sydney, New South Wales, Australia. [2]ARC Centre of Excellence for Climate Extremes, University of New South Wales, Sydney, New South Wales, Australia. [3]Science and Innovation Group, Australian Bureau of Meteorology, Melbourne, Victoria, Australia.
*Corresponding author. Email: h.ayat@unswalumni.com

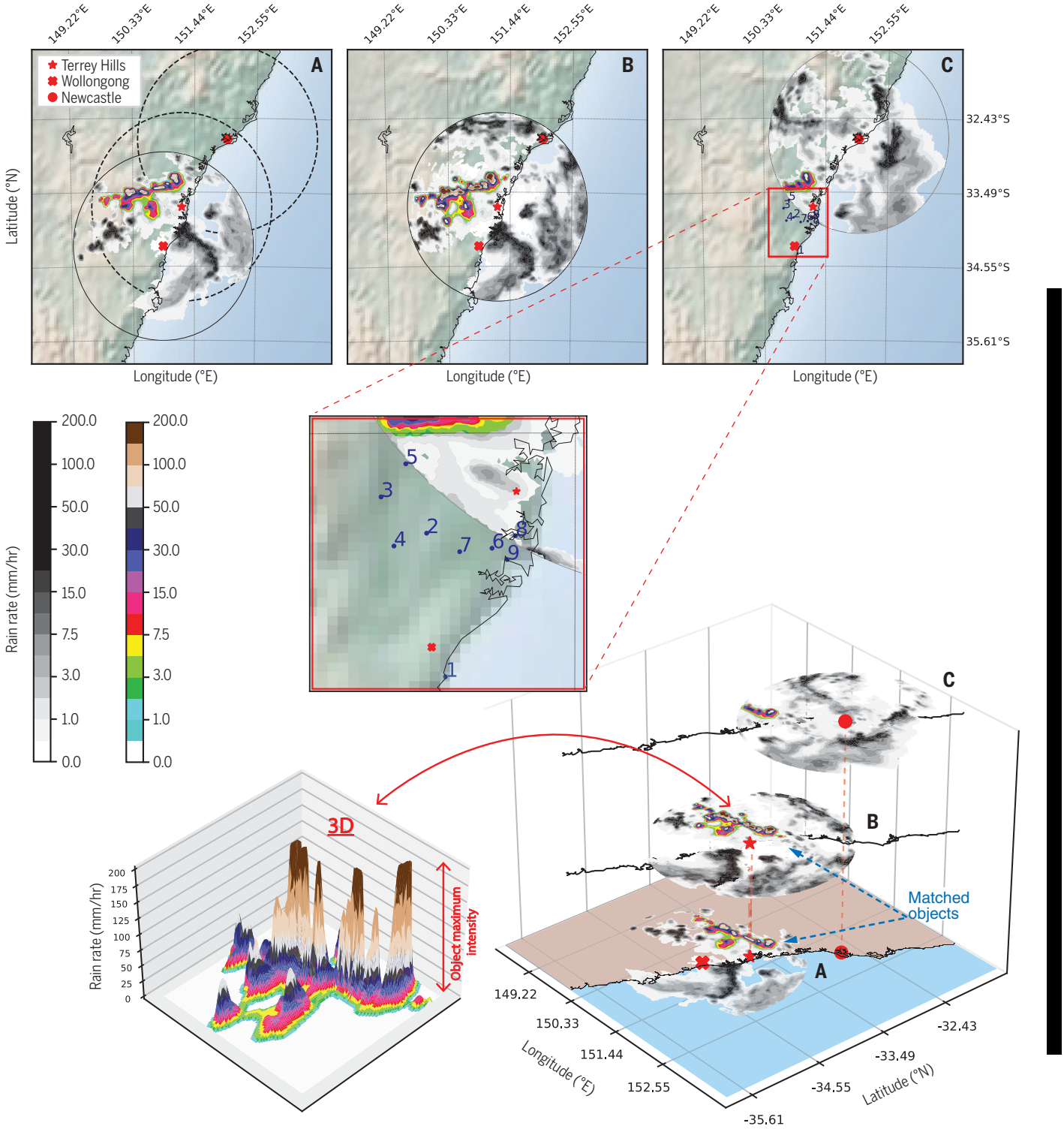


Fig. 1. The locations of the selected weather radar and AWS rain gauges that captured an event on 16 December 2015 at 02:06:00 UTC. (**A** to **C**) Fields of view of the Wollongong (A), Terrey Hills (B), and Newcastle (C) radars while observing the event. In (A), the bottom solid circle shows the Wollongong radar field of view, and the middle and top dashed circles show the relative locations of the Terrey Hills and Newcastle radar fields of view, which are shown in (B) and (C). The highlighted rainfall patterns are objects that were detected using the object-detection technique. The detected objects in the Wollongong (A) and Terrey Hills (B) radars are matched with the total interest value of 1 (see supplementary methods for more details). The numbers in the magnified view from (C) represent the AWS rain gauges. An example of the maximum intensity of an object is shown in the 3D view of the detected object in Terrey Hills radar (B) at the bottom left. An example of a matched object is shown at the bottom right.

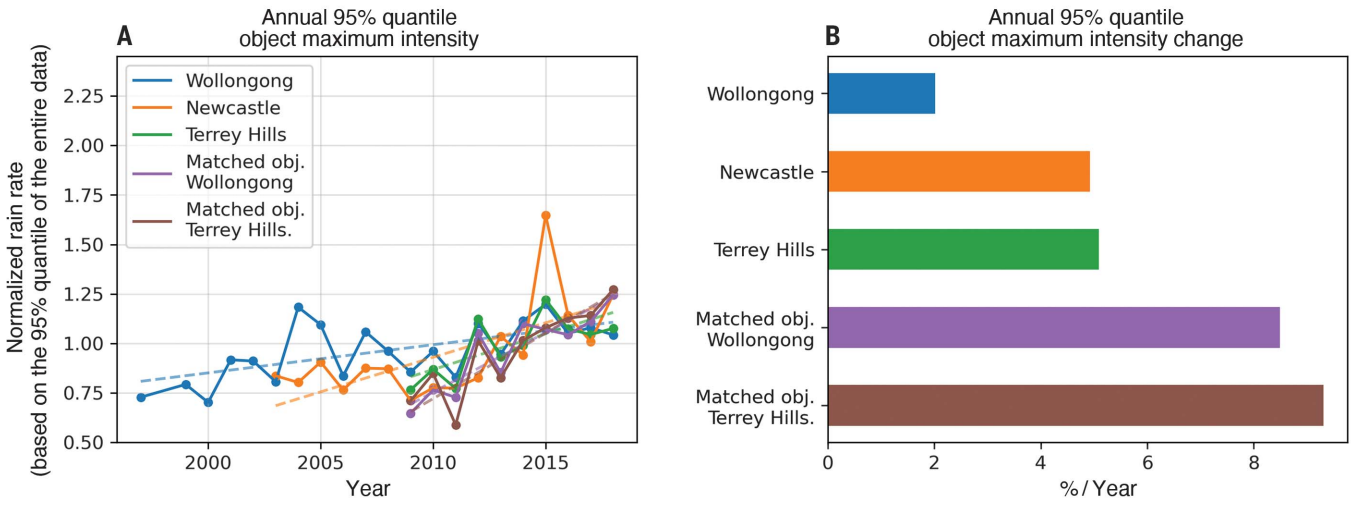


Fig. 2. Annual 95% quantile of the object maximum intensity trend in the radar datasets. (**A**) The 95% quantile time series of annual rainfall. The dashed lines are the 95% quantile regression fitted over the data matched with their years of occurrence. The values on the *y* axis are normalized based on the 95% quantile of the entire dataset of each radar. (**B**) Calculated slope of changes using the quantile regression analysis. All the calculated trends are significant at the level of 0.05. Blue, Wollongong radar; orange, Newcastle radar; green, Terrey Hills radar; purple, objects in Terrey Hills that are matched with their Wollongong counterparts; brown, objects in Wollongong that are matched with their Terrey Hills counterparts.

hardware and location on the calculated trends in the previous step. For this purpose, we used an object-matching technique to isolate the objects that are observed in two radars at each time step. By using this technique, an object from the Terrey Hills radar dataset is matched with its counterpart in the Wollongong radar dataset at each time step if they have similar properties (here location and area; see supplementary methods for more details).

In the next step, all rainfall data from 1999 to 2017 (including wet and dry data) from nine high-resolution (1 min) automatic weather stations (AWSs) are resampled to subhourly (10 min), hourly, and daily time scales. The 99 and 99.9% quantile regressions are fitted over the data from each year to investigate the trend of rainfall extremes from a point-based perspective. We repeated the same procedure for the same period over Wollongong radar data and sampled at the locations of AWS rain gauges to directly compare the radar to other platforms. Note that the 95% quantile is targeting low rainfall intensities in the point-based approach because wet and dry records were included (see supplementary methods). Hence, we have selected 99 and 99.9% quantiles to target the extremes. Note that to compare the rate of change between the radars and gauge stations and estimate a rate of change independent of an observational platform, we normalized the results by dividing them by the average of the data for each radar and gauge station.

Figure 2 presents the time series and linear trends of the annual extreme (95% quantile across all objects) maximum rain rate seen in the object for each radar separately and for the matched objects in the Terrey Hills and Wollongong radars. There are positive trends in all radar datasets that are statistically significant at the 0.05 level. The radars thus corroborate each other and suggest an upward trend of at least 2% per year since 1997 (accelerating to up to 8% more recently) in the maximum intensity. The actual trends look strongest for the most extreme values, and any approach or instrument that is reaching farther out on the distribution tail encounters stronger trends. For example, trends are even stronger, and very similar, in the matched objects in both the Wollongong and Terrey Hills datasets. This is related to the matching process that often excludes small-scale nonextreme storms with shallow vertical structures because a pair of matched objects is detected at different distances from the radars and different altitudes (fig. S6) and often includes well-developed and more extreme storms. Similar results have also been found using the 90% quantile of data (see fig. S1).

Figure 3 shows the detected trends of subhourly (10 min), hourly, and daily extreme (99%) and very extreme (99.9%) rainfall from a point-based perspective using nine AWS rain gauges (orange) and the Wollongong radar data (blue) at the locations of the AWS rain gauges by applying the quantile regression technique. The trends that are not statistically significant are shown in gray. Figure 3A indicates that most AWS rain gauges (six out of nine) show robust upward trends in very extreme rainfall (99.9%) at a subhourly resolution (10 min) with a rate of up to 2% per year. These trends are even more clear in the Wollongong radar data, because all the selected pixels (at the AWS locations) are showing statistically significant positive trends. These 10-min trend results are in agreement with what we have observed in the trend of object maximum intensity. However, fewer stations (both in radar and AWS data) show robust positive trends in hourly accumulations, especially in AWS data (Fig. 3B). Note that a daily trend analysis for the 99.9% quantile is not possible because the maximum number of available daily data in a year is less than 1000. The observed trends in the extreme rainfall (99%) are weaker in subhourly accumulations (Fig. 3C) but stronger and clearer in hourly accumulations (Fig. 3D). However, this signal becomes less clear in daily accumulations, for which most of the stations do not show robust upward trends (Fig. 3E). Note that a possible reason for steeper observed trends in radars is that radars observe higher-level precipitation, and when the storms get stronger, they may be lofting precipitation higher, so the fraction reaching the ground might decrease a bit, causing the higher trend. Conversely, gauge records are sensitive to winds, which are more severe during heavy rainfall and cause gauge undercatch, possibly resulting in an underestimation of trends in gauge instruments. This may explain all or some of the differences in trend magnitudes.

We also investigated the relationships between climate indices [i.e., El Niño–Southern Oscillation (ENSO), Indian Ocean Dipole (IOD), and Southern Annular Mode (SAM)] and object maximum intensity in another study (*46*) (using Wollongong radar data and a multivariate approach), and the results did not show any link between intensity and any of the modes in this

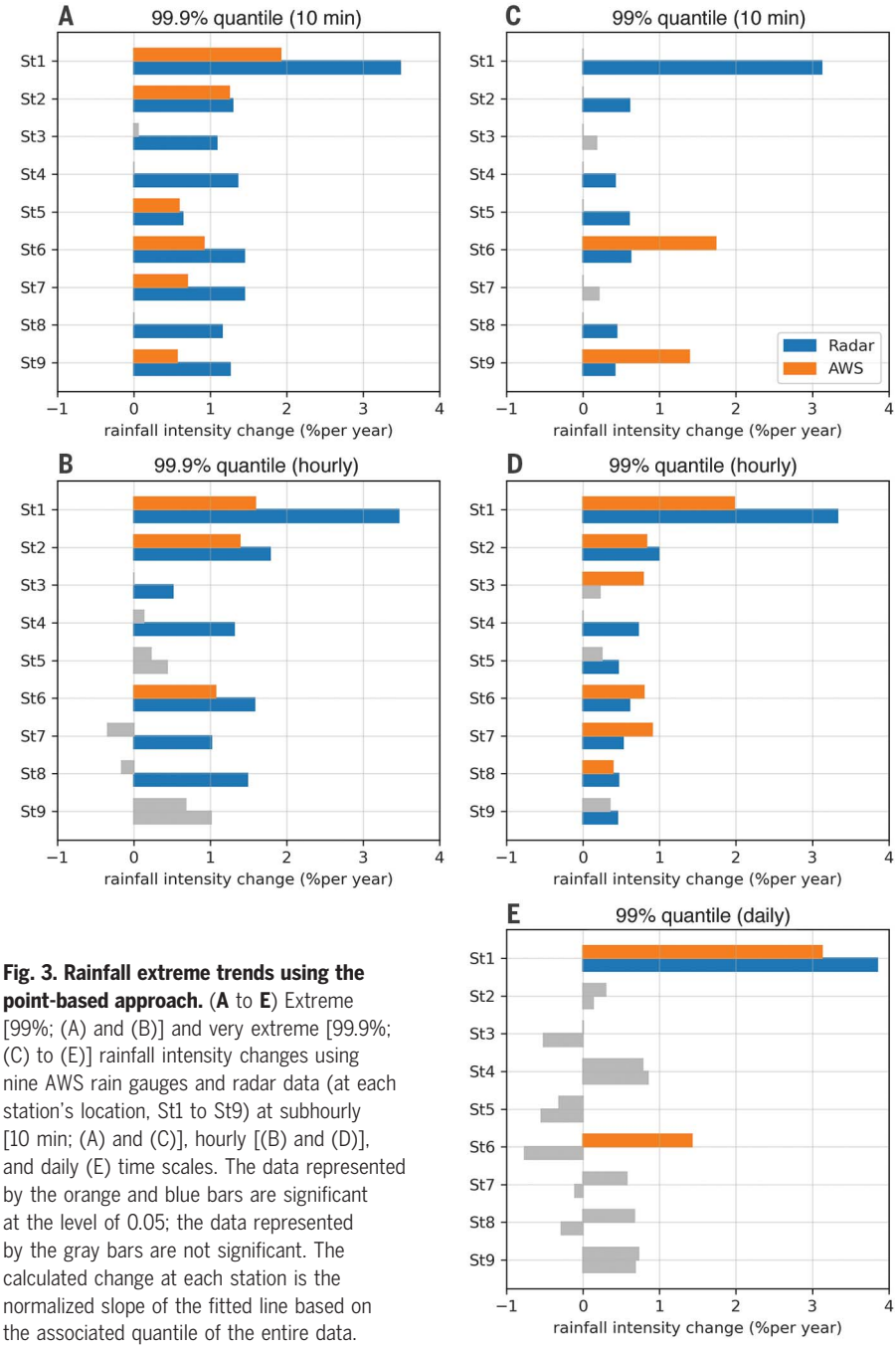


Fig. 3. Rainfall extreme trends using the point-based approach. (**A** to **E**) Extreme [99%; (A) and (B)] and very extreme [99.9%; (C) to (E)] rainfall intensity changes using nine AWS rain gauges and radar data (at each station's location, St1 to St9) at subhourly [10 min; (A) and (C)], hourly [(B) and (D)], and daily (E) time scales. The data represented by the orange and blue bars are significant at the level of 0.05; the data represented by the gray bars are not significant. The calculated change at each station is the normalized slope of the fitted line based on the associated quantile of the entire data.

region. Similar findings were also reported using point-based approaches for short-duration accumulations of (hourly and/or subhourly) rainfall extremes in Australia (*1*) and eastern Australia (*2*), suggesting that these trends are not a part of natural variability.

The high rate of change found here (at least 20% per decade) has not been reported in previous studies in this region or other parts of the world. Although we find positive trends at gauge stations as examined in previous studies, the rates of change, except at one station, are much smaller than 20% per decade. This shows that the trends in gauge stations are much noisier and more uncertain than those in weather radars because so many storms are missed or unrepresentatively sampled by the gauges (*36*).

It is interesting that the observed positive trends in subhourly rainfall extremes gradually decline at longer accumulation intervals, suggesting that even in places with little trend in daily rainfall extremes, there may still be increasing risks of flash floods due to the intensification of subhourly rainfall extremes. This result is consistent with Kendon *et al.* (*40*), who showed that changes in 10-min and hourly precipitation emerge before changes in daily precipitation in a CPM simulation over the southern United Kingdom. Thus, there is observational evidence that suggests that the surprising trend reported here may be occurring in many parts of the world, and it is qualitatively consistent with model predictions in warmer atmospheres (*37*, *38*).

Although understanding the changes in rainfall extremes globally has been inhibited by a lack of data (*4*), this study shows that weather radar data could be a valuable means toward overcoming the existing limitations by using the analysis framework introduced here. Better spatial coverage and resolution of weather radar data provide potential opportunities to study storm trends using more details such as changes in different storm types and properties other than intensity and frequency (i.e., size, shape, translation speed, and so on). In addition, by using the object-based approach and object-matching technique presented in this study, CPMs can be better evaluated through simulations of recent trends so that there is a higher confidence in their future projections.

REFERENCES AND NOTES

1. S. B. Guerreiro *et al.*, *Nat. Clim. Chang.* **8**, 803–807 (2018).
2. O. U. Laz, A. Rahman, A. Yilmaz, K. Haddad, *J. Water Clim. Chang.* **5**, 667–675 (2014).
3. K. E. Trenberth, *Clim. Change* **39**, 667–694 (1998).
4. H. J. Fowler *et al.*, *Nat. Rev. Earth Environ.* **2**, 107–122 (2021).
5. D. R. Archer, H. J. Fowler, *J. Flood Risk Manag.* **11**, S121–S133 (2018).
6. H. J. Fowler *et al.*, *Philos. Trans. R. Soc. London Ser. A* **379**, 20190542 (2021).
7. E. Hajani, A. Rahman, E. Ishak, *Hydrol. Sci. J.* **62**, 2160–2174 (2017).
8. S. Westra *et al.*, *Rev. Geophys.* **52**, 522–555 (2014).
9. E. M. Fischer, R. Knutti, *Nat. Clim. Chang.* **6**, 986–991 (2016).
10. S. C. Scherrer *et al.*, *J. Geophys. Res. Atmos.* **121**, 2626–2637 (2016).
11. J. Rajczak, C. Schär, *J. Geophys. Res. Atmos.* **122**, 10773–10800 (2017).
12. T. Schneider, P. A. O'Gorman, X. J. Levine, *Rev. Geophys.* **48**, RG3001 (2010).
13. C. Wasko, A. Sharma, F. Johnson, *Geophys. Res. Lett.* **42**, 8783–8790 (2015).
14. P. Berg, C. Moseley, J. O. Haerter, *Nat. Geosci.* **6**, 181–185 (2013).
15. G. Lenderink, E. van Meijgaard, *Environ. Res. Lett.* **5**, 025208 (2010).
16. G. Lenderink, E. van Meijgaard, *Nat. Geosci.* **1**, 511–514 (2008).
17. K. Lochbihler, G. Lenderink, A. P. Siebesma, *Geophys. Res. Lett.* **44**, 8629–8636 (2017).
18. I.-H. Park, S.-K. Min, *J. Clim.* **30**, 9527–9537 (2017).
19. G. Lenderink, H. Y. Mok, T. C. Lee, G. J. van Oldenborgh, *Hydrol. Earth Syst. Sci.* **15**, 3033–3041 (2011).
20. Z. Li, P. A. O'Gorman, *J. Clim.* **33**, 7125–7139 (2020).
21. S. Pfahl, P. A. O'Gorman, E. M. Fischer, *Nat. Clim. Chang.* **7**, 423–427 (2017).
22. R. Barbero, H. J. Fowler, G. Lenderink, S. Blenkinsop, *Geophys. Res. Lett.* **44**, 974–983 (2017).
23. C. Clarke, M. Hulley, J. Marsalek, E. Watt, *Can. J. Civ. Eng.* **38**, 1175–1184 (2011).
24. Y.-R. Chen, B. Yu, G. Jenkins, *J. Hydrometeorol.* **14**, 1356–1363 (2013).

25. S. M. Herath, R. Sarukkalige, V. T. V. Nguyen, *J. Hydrol.* **556**, 1171–1181 (2018).
26. D. Jakob, D. J. Karoly, A. Seed, *Nat. Hazards Earth Syst. Sci.* **11**, 2263–2271 (2011).
27. D. Jakob, D. J. Karoly, A. Seed, *Nat. Hazards Earth Syst. Sci.* **11**, 2273–2284 (2011).
28. S. Westra, S. A. Sisson, *J. Hydrol.* **406**, 119–128 (2011).
29. A. G. Yilmaz, B. J. C. Perera, *J. Hydrol. Eng.* **19**, 1160–1172 (2014).
30. F. Zheng, S. Westra, M. Leonard, *Nat. Clim. Chang.* **5**, 389–390 (2015).
31. K. Arnbjerg-Nielsen, *Water Sci. Technol.* **54**, 1–8 (2006).
32. E. Arnone, D. Pumo, F. Viola, L. V. Noto, G. La Loggia, *Hydrol. Earth Syst. Sci.* **17**, 2449–2458 (2013).
33. H. Madsen, K. Arnbjerg-Nielsen, P. S. Mikkelsen, *Atmos. Res.* **92**, 343–349 (2009).
34. A. H. Syafrina, M. D. Zalina, L. Juneng, *Theor. Appl. Climatol.* **120**, 259–285 (2015).
35. C. Xiao, P. Wu, L. Zhang, L. Song, *Sci. Rep.* **6**, 38506 (2016).
36. K. Lengfeld *et al*., *Environ. Res. Lett.* **15**, 085003 (2020).
37. C. Li *et al*., *Geophys. Res. Lett.* **46**, 6885–6891 (2019).
38. J. Olsson, K. Foster, *Nord. Hydrol.* **45**, 479–489 (2013).
39. S. C. Chan, E. J. Kendon, N. M. Roberts, H. J. Fowler, S. Blenkinsop, *Environ. Res. Lett.* **11**, 094024 (2016).
40. E. J. Kendon, S. Blenkinsop, H. J. Fowler, *J. Clim.* **31**, 2945–2964 (2018).
41. L. V. Alexander *et al*., *Environ. Res. Lett.* **14**, 125008 (2019).
42. F. B. Avila *et al*., *Weather Clim. Extrem.* **9**, 6–16 (2015).
43. H. Ali, N. Peleg, H. J. Fowler, *Geophys. Res. Lett.* **48**, e2021GL093798 (2021).
44. H. Ayat, J. P. Evans, A. Behrangi, *Remote Sens. Environ.* **259**, 112417 (2021).
45. J. Soderholm, V. Louf, A. Protat, R. A. Warren, Australian Operational Weather Radar Level 2 Dataset, NCI Data Catalogue (2022); https://doi.org/10.25914/40KE-NM05.
46. H. Ayat, J. P. Evans, S. C. Sherwood, J. Soderholm, Research Square [Preprint] (2021); https://doi.org/10.21203/rs.3.rs-1134660/v1.
47. H. Ayat, J. P. Evans, S. Sherwood, A. Behrangi, *J. Hydrometeorol.* **22**, 43–62 (2021).

ACKNOWLEDGMENTS

This research was undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian government. **Funding:** This work was supported by the Australian Research Council as part of the Center of Excellence for Climate Extremes (grant CE170100023 to J.P.E. and S.C.S.). **Author contributions:** Conceptualization: H.A., J.P.E., S.C.S.; Methodology: H.A., J.P.E., S.C.S.; Data analysis: H.A.; Investigation: H.A., J.P.E., S.C.S., J.S.; Supervision: J.P.E., S.C.S., J.S.; Writing – original draft: H.A.; Writing – review and editing: J.P.E., S.C.S., J.S. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Radar data used in this study are available at https://geonetwork.nci.org.au/geonetwork/srv/eng/catalog.search#/metadata/f8188_7912_6774_5057. AWS data are also accessible from the following link: http://www.bom.gov.au/climate/data/stations/. The source code of the object-based method used in this study is available at https://github.com/H0omaN/MTD-Modified.

License information: Copyright © 2022 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abn8657
Materials and Methods
Figs. S1 to S10
Table S1
References (*48*–*50*)

Submitted 29 December 2021; accepted 12 October 2022
10.1126/science.abn8657

METALLURGY

Inhibiting creep in nanograined alloys with stable grain boundary networks

B. B. Zhang[1], Y. G. Tang[1,2], Q. S. Mei[3], X. Y. Li[1]*, K. Lu[1,4]*

Creep, the time-dependent deformation of materials stressed below the yield strength, is responsible for a great number of component failures at high temperatures. Because grain boundaries (GBs) in materials usually facilitate diffusional processes in creep, eliminating GBs is a primary approach to resisting high-temperature creep in metals, such as in single-crystal superalloy turbo blades. We report a different strategy to inhibiting creep by use of stable GB networks. Plastic deformation triggered structural relaxation of high-density GBs in nanograined single-phased nickel-cobalt-chromium alloys, forming networks of stable GBs interlocked with abundant twin boundaries. The stable GB networks effectively inhibit diffusional creep processes at high temperatures. We obtained an unprecedented creep resistance, with creep rates of ~10^{-7} per second under gigapascal stress at 700°C (~61% melting point), outperforming that of conventional superalloys.

When a metal is stressed far below its yield strength at elevated temperatures, a continuously increasing strain is often induced—a process known as creep. A great number of failures of materials and components serving at high temperatures are attributed to creep or to its combination with other degradation processes (*1*), costing billions of dollars annually for the repair and replacement of parts in advanced devices. The ever-growing demand of higher fuel efficiency and reliability of turbines, nuclear reactors, and devices in chemical industries calls for a constant improvement of high-temperature creep resistance in advanced alloys (*1*).

Creep is a time-dependent deformation process controlled by atomic diffusion and dislocation glide. To resist creep in materials, one may increase their high-temperature strength, minimize atomic diffusion, or both. Alloying may strengthen materials at high temperatures by solution hardening with heat-resistant elements tungsten (W), molybdenum (Mo), and rhenium (Re) in superalloys (*2*) or by forming more stable phases for strengthening or for pinning grain boundary (GB) motion at high temperatures (*3*). For example, the superalloys that serve at temperatures higher than 650°C are heavily alloyed, in which the strengthening γ'/γ'' phases constitute up to 60% in volume.

GBs in materials are usually regarded to be deleterious to creep. Atomic diffusivity along GBs in metals is several orders of magnitude higher than that in lattices above 0.5 T_m (T_m is the melting point in kelvin), aggravating the GB diffusional (Coble) creep (*4*). GBs facilitate vacancy migration and deformation of grains in the Nabarro-Herring creep (*5*, *6*). In addition, the strengthening effect of GBs, although pronounced in metals at room temperature (RT), vanishes typically above 0.5 T_m, at which GB processes (migration or sliding) become prominent (*7*). Hence, eliminating GBs in materials is another primary approach to minimizing atomic diffusion and resisting creep at high temperatures, as practiced in producing turbo blades of single-crystal or directional solidified superalloys (*2*). But the creep resistance of the alloys achieved so far are moderate. The creep-stress levels at >0.5 T_m are lower than 0.8 GPa even for the most heavily alloyed single-crystal superalloy (CMSX-4) (*2*).

Contradictory to the conventional wisdom of eliminating GBs, we proposed introducing abundant GBs to form stable GB networks in metals to inhibit creep by means of suppressing atomic diffusion and hardening at high temperatures, simultaneously. The idea was inspired by recent studies on GB relaxation in metals—namely, that GBs can adjust their structures into low-energy states through interactions with partial dislocations as triggered by plastic straining in a number of metals and alloys with nano-sized grains (*8*). The relaxed GBs become more stable against migration under thermal (*8*) and mechanical (*9*) activation, respectively, implying that atomic diffusion associated with the relaxed GBs are retarded at elevated temperatures. In addition, the strengthening effect of the stabilized GBs may remain at higher temperatures. Both features may resist creep deformation, as long as the GBs are capable of inhibiting diffusion and remaining strong under a combined stimuli with high stress and high temperatures.

We performed an experimental study by taking a single-phased nickel-cobalt-chromium (NiCoCr) alloy as an example. We found that the diffusional creep processes is effectively suppressed in the nanograined alloy with the stable GB networks. We obtained a high creep resistance with a steady-state creep rate of ~10^{-7} s^{-1} under gigapascal stress at 700°C (0.61 T_m), outperforming that of the conventional superalloys.

We used a commercial single-phased NiCoCr alloy with a chemical composition in weight percentage of 34.1Ni–33.9Co–20.9Cr–10.2Mo–0.9Ti (or called MP35N, a composition similar to a medium-entropy alloy). The alloy has a stable face-centered cubic structure without any precipitation and phase transformation during deformation and thermal treatments (*10*).

According to our proposed mechanism, GB relaxation can be triggered by plastic straining with grains smaller than a critical size that corresponds to a transition in deformation mechanism from full dislocation slip to partials (*8*). We estimated the critical grain size in the alloy, following the classical dislocation theory, to be about 37 nm (*11*). This suggests that plastic straining of the alloy with grains smaller than 37 nm may induce structure relaxation of GBs into low-energy configurations. Hence, we processed the alloy bars (with an initial structure as in fig. S1) by using a surface mechanical grinding treatment (the processing parameters in table S1) for refining grains into the desired size regime in the surface layer, a process applied previously in many other metals and alloys (*8*).

After the treatment, uniformly distributed equiaxed nanograins formed the top surface layer, about 25 μm thick in the treated samples (Fig. 1A). The grain sizes ranged from few nanometers to about 30 nm (average, 9 ± 3 nm; error is standard deviation) (Fig. 2B) under transmission electron microscopy (TEM) measurements, which is well below the critical size. The grain size is roughly homogeneous along depth within this layer. Homogeneous composition distributions were detected for each element without any difference across GBs in high-magnification energy-dispersive x-ray spectroscopy (EDS) mappings (Fig. 1B). Segregation of elements or second-phase particles at GBs is absent.

Under high-resolution TEM (HRTEM), we found that nanograins are typically separated by faceted boundaries, with segments of {111} atomic planes connected by steps (Fig. 1C). Statistical measurements along the [110] axis showed that the number fraction of nanograins of which the majority of boundaries (>50%) are faceted {111} is about 70%. Procession electron diffraction (PED) statistical measurements under TEM (Fig. 1, D and E) revealed a random distribution of grain orientations with a considerable fraction of low ΣCSL (coincidence-site lattice) boundaries (<Σ30) (44%). These are typical structural features of GBs in relaxed states with low excess energies, analogous to

[1]Shenyang National Laboratory for Materials Science, Institute of Metal Research, Chinese Academy of Sciences, Shenyang 110016, China. [2]School of Materials Science and Engineering, University of Science and Technology of China, Shenyang 110016, China. [3]School of Power and Mechanical Engineering, Wuhan University, Wuhan 430072, China. [4]Liaoning Academy of Materials, Shenyang 110004, China.
*Corresponding author. Email: xyli@imr.ac.cn (X.Y.L.); lu@imr.ac.cn (K.L.)

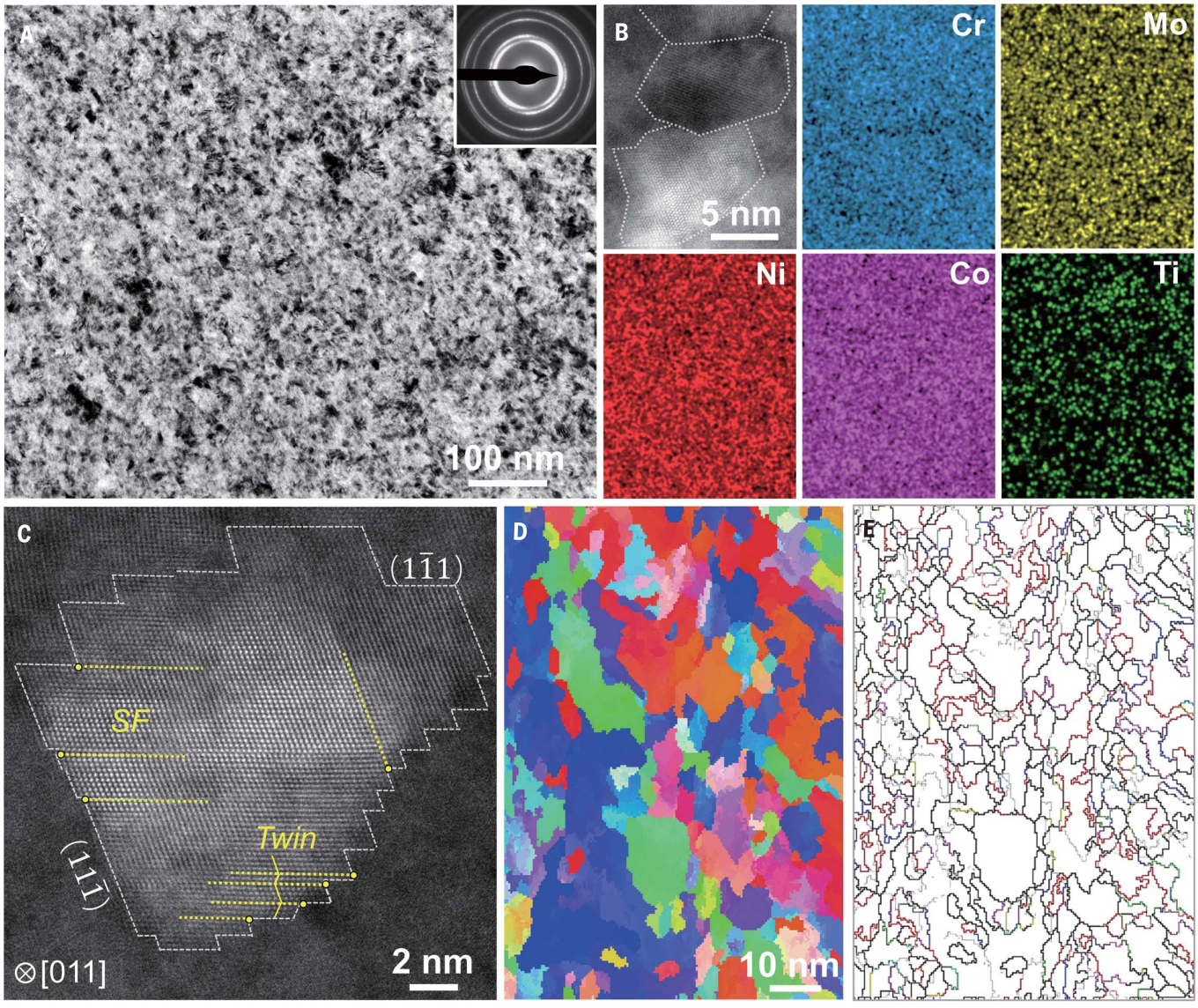


Fig. 1. Structures of the nanograined NiCoCr alloy with stable GBs. (**A**) A bright-field TEM image showing extremely fine grains. (**B**) A high-angle annular dark-field (HAADF) image of several grains, outlined by dotted lines, with corresponding EDS mappings of different elements. (**C**) An HRTEM image of an individual grain with faceted GBs, which contains twins and stacking faults (SFs). (**D**) An inverse pole figure (IPF) image with (**E**) a distribution map of GB characters from precession electron diffraction analysis. Colored lines indicate different boundaries: red, twin boundaries; gray, low-angle GBs (5 to ~15°); black, ordinary high-angle GBs (>15°); and other colors, other special boundaries ($\Sigma < 30$).

that in copper (Cu) and Ni with relaxed GBs (*8*). They are distinct from the "high-energy nonequilibrium" GBs in metals that have undergone severe plastic deformation (*12*). Copious $\Sigma 3$ twin boundaries and stacking faults were detected in the nanograins (Fig. 1C), with a number density of twin- and stacking fault–GB triple line of ~0.13 nm^{-1}, supporting that partial dislocation activities dominated the deformation and refinement processes as expected. Hence, we believe that a large population of GBs were structurally relaxed into more stable states triggered by the mechanical processing, forming dense GB networks of abundant stable GBs interwoven with copious in-grain twin boundaries and stacking faults in this nanograined sample (denoted as SNG-9).

For comparison, we cut out another set of specimens in the subsurface layer from the same as-processed bar in which grains were larger than the critical size. A foil of about 25 μm thick consists of roughly equiaxed and randomly oriented grains with an average transverse size of 42 ± 12 nm (aspect ratio of ~1.6) (denoted as the NG-42 sample) (fig. S2), which is similar to the nanograined structures in metals from severe plastic deformation (*12*). Most GBs are ordinary with curved morphologies, with only a very small fraction of grains having faceted boundaries. Twins and stacking faults were rarely seen with a much lower density of twin- and stacking fault–GB triple line (~0.02 nm^{-1}) than that of SNG-9. These suggest that GB relaxation was not induced in this sample, which has a chemical composition identical to that of the SNG-9 sample.

We characterized thermal stability of the samples by annealing at elevated temperatures. In NG-42, obvious grain coarsening to the submicro-scale was detected as annealed at >700°C (Fig. 2A and fig. S3), which is analogous to the recrystallization in the coarse-grained (CG) MP35N sample. However, nanograins in SNG-9 remain stable in

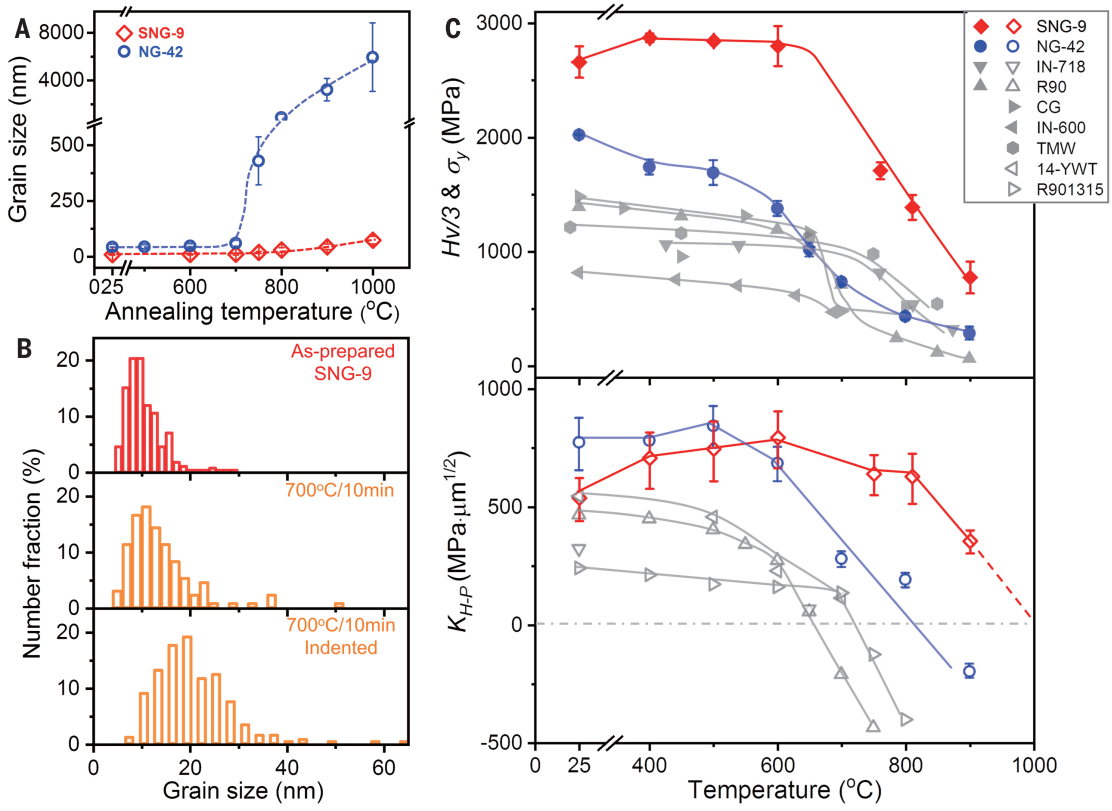


Fig. 2. Thermal stability and temperature dependence of strength and GB strengthening. (**A**) Variation of grain size as a function of annealing temperature. (**B**) Grain size distributions of the as-prepared SNG-9 sample, the sample annealed at 700°C for 10 min, and that underneath the indents after creep tests. (**C**) Yield strength and GB-strengthening coefficient *K* versus temperature for SNG-9 and NG-42 samples, in comparison with several typical superalloys, including Ni-based superalloys [TMW (*16*) and Inconel 600 and 718 (*17*, *21*)] and oxide dispersion–strengthened (ODS) Ni–based superalloys [R90 and R901315 (*15*)], ODS-steel [14-YWT (*20*)], and the deformed CG MP35N alloy (*18*).

morphology and size after annealing at 700°C. Grain sizes increase slightly to ~30 ± 10 nm after annealing at 800°C for 12 hours, and below 100 nm at 1000°C (Fig. 2A and figs. S3 and S4). The higher thermal stability of the SNG-9 is contradictory to the normal "smaller, less-stable" trend (*13*) but in alignment with that in nanograined Cu, Ni, and aluminium-magnesium (Al–Mg) samples with relaxed GBs (*8*, *14*). The inherent stability against grain coarsening can be reasonably attributed to the stable GB networks consisting of abundant relaxed GBs interlocked with copious twin boundaries and stacking faults.

We measured Vickers hardness (H_V) of the samples at various temperatures from RT to 900°C. Yield strength (σ_y) was converted with a Tabor factor of 3.0 (Fig. 2C) (*1*). The strength of the NG-42 sample is around 2 GPa at RT and decreases at higher temperatures and pronouncedly at >600°C, which is similar to that in conventional superalloys (*15*–*18*). The SNG-9 sample exhibits a yield strength of 2.6 GPa at RT and increases slightly to 2.8 GPa at 400° to 600°C. Strength decreases above 600°C, although the values are well above that of NG-42 and other alloys over the entire temperature range. We plotted the measured RT strength values of NG-42 and SNG-9 samples and found that they fall nicely on the Hall-Petch line extrapolated from the CG counterparts, which is indicative of effective GB strengthening at RT in the SNG-9 sample.

The GB strengthening capability at different temperatures (*T*) can be evaluated by using the coefficient *K* in the Hall-Petch relationship (*19*), $H_V = H_{V0} + (E_T/E_{RT})^{1/2} \cdot K \cdot d^{-1/2}$, where H_{V0} is intragranular hardness that can be obtained from the CG samples; *d* is the average grain size; and E_{RT} and E_T are the elastic modulus at RT and *T*, respectively (*11*). For the NG-42 sample, *K* decreases above 500°C, with an equicohesive temperature (when $K = 0$, GB strengthening vanishes) of ~800°C. In SNG-9, *K* clearly decreases at >810°C, with an extrapolated equicohesive temperature of ~1000°C, which is much higher than that of NG-42 and other alloys (*15*, *20*, *21*). These observations suggest that the relaxed GBs remain structurally stable and gain strength at high temperatures of up to ~1000°C.

We evaluated creep behaviors of the SNG-9 sample in a temperature range of 600° to 750°C by using the indentation-creep method (*22*) on a Nano Test Vantage (Micro Materials, UK) equipped with a cubic boron nitride Berkevich indenter. For comparison, we measured sample NG-42 and a deformed CG sample (with grain sizes of ~50 μm) under the same conditions. The steady-state creep stress and the corresponding strain rate were derived from the indentation displacement-time curves (Fig. 3A), with a holding time exceeding 300 s.

For the CG sample, creep rates at 700°C ranged from 5.6×10^{-4} to 3.2×10^{-4} s^{-1} under a stress of 301 to 239 MPa (~0.4% *G*; *G*, shear modulus), which is in agreement with the values calculated from the dislocation creep mechanism as in the Weertman-Ashby deformation map (Fig. 3B). The NG-42 sample exhibits creep rates of 5.6×10^{-4} to 3.9×10^{-4} s^{-1} under a stress of 100 to 77 MPa (~0.1% *G*) at 700°C, and 4.3×10^{-4} to 1.2×10^{-4} s^{-1} under 832 to 720 MPa (~1.1% *G*) at 600°C, respectively. The calculated stress exponent is about 1.3, which is in agreement with the diffusional creep process. Apparently, the creep resistance is lower in NG-42 because of its much higher GB density relative to the CG sample. Considering the concurrent grain coarsening induced by the indentation loading (average grain sizes increase from an initial ~59 to ~400 nm at 700°C), we calculated the creep rates of GB diffusional (Coble) and GB sliding mechanisms following equations in (*11*) for sample NG-42 under a stress level of ~0.1% *G* at 700°C. We found that the two calculated rates are comparable; both are close to the measured rates (for example, Coble creep rates in Fig. 3B). It suggests that both mechanisms may be operative during the creep process, which is consistent with that observed in other nanograined metals and alloys (*23*). An obvious decrease in texture intensity was detected in the sample after creep tests (fig. S6), verifying the GB sliding mechanism that may randomize the grain orientation distributions.

Very low creep rates were observed in the SNG-9 sample, with a high strength at elevated

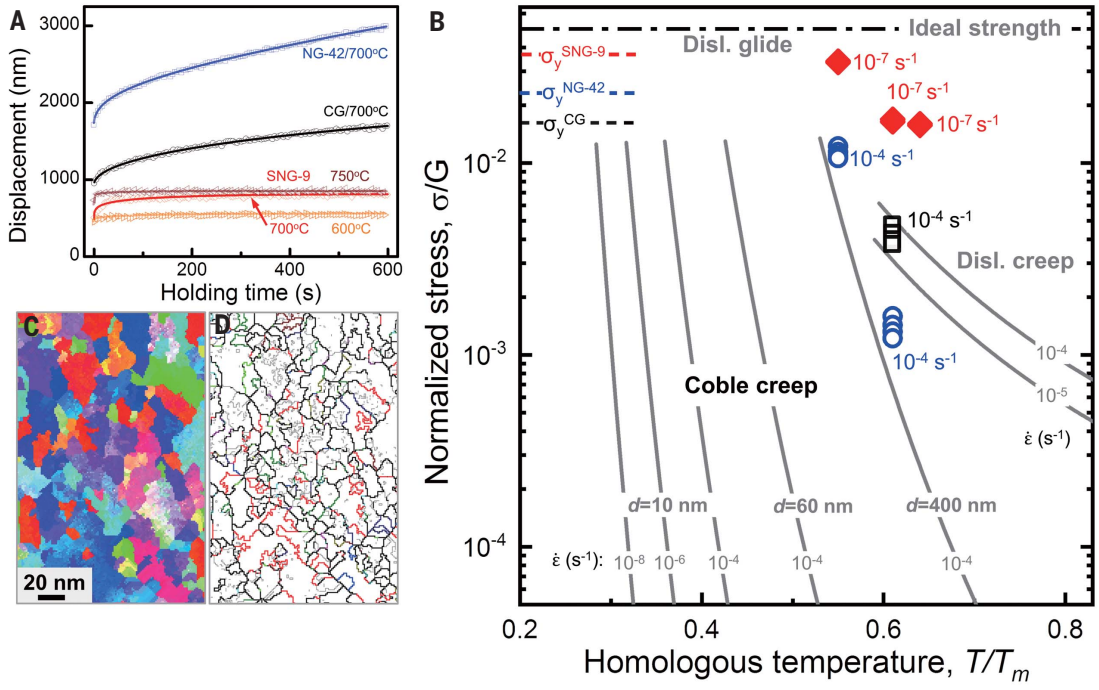


Fig. 3. Indentation creep response and mechanism. (**A**) Displacement versus time curves for the SNG-9, NG-42, and deformed CG samples in the holding stage of indentation creep at various temperatures. (**B**) Weertman-Ashby map of deformation mechanism for the NiCoCr alloy, with theoretical constant Coble creep rates for different grain sizes as well as dislocation creep rates. The measured creep data for the SNG-9, NG-42, and CG samples are overplotted for comparison. (**C**) A typical IPF image with (**D**) a distribution map of GB characters underneath the indents after creep tests at 700°C in the SNG-9 sample.

temperatures. The steady-state creep rates were 9.8×10^{-8} to 8.5×10^{-7} s^{-1} under a stress of 2274 MPa (3.3% *G*) at 600°C, 5.2×10^{-8} to 9.5×10^{-7} s^{-1} under 1036 MPa (1.6% *G*) at 700°C, and (2.4 to 9.2) $\times 10^{-7}$ s^{-1} under 959 MPa at 750°C, respectively. Very large stress exponents (>100) were obtained at these temperatures, implying a very high creep resistance. At 700°C, the creep stress is more than 10 times higher than that in NG-42, whereas the strain rates are three orders of magnitude smaller. The creep resistance is much higher than that of NG-42 and the deformed CG samples, which is contradictory to the conventional wisdom that smaller grains lower the creep resistance. Our measured creep data (Fig. 3B) are far apart from the calculated values following the Coble creep mechanism with a grain size of 10 to 20 nm (grain coarsening during the indentation is marginal) (Figs. 2B and 3C). The observed strain rates of SNG-9 are several orders of magnitude smaller than that from the Nabarro-Herring creep mechanism. Clearly, the diffusional creep processes are suppressed in the SNG-9 sample below 750°C, at least.

We detected a marginal microstructure change underneath the indents in the SNG-9 sample after the creep tests. Grain sizes increased slightly from an average value of 12 ± 4 nm to 19 ± 6 nm after the indentation at 700°C (Figs. 2B and 3C). Similar to the as-prepared structure, randomly oriented nanograins with abundant low-energy GBs including twin boundaries and faults were seen under HRTEM (Fig. 3, C and D). The texture intensity was essentially unchanged (fig. S5). These observations show that both GB migration and GB sliding are negligible in the creep deformation, in contrast to that in the NG-42 sample (fig. S6) and other nanograined metals (*24*). From HRTEM observations of many grains in the as-prepared SNG-9 samples, we did not find dislocations in the interiors of the extremely fine grains. The dislocation starvation, which is consistent with the documented literature (*25*), may play a role in the observed suppression of GB migration and GB sliding processes. Suppressed GB activities in SNG-9 in creep deformation explains the measured creep properties and the hot hardness results,

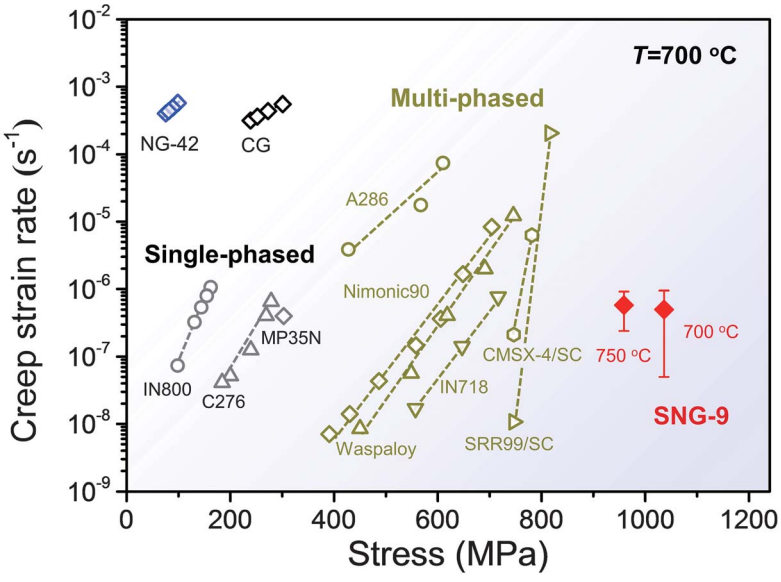


Fig. 4. Creep resistance of various alloys at 700°C. A plot of creep rate versus stress for the SNG-9, NG-42 and deformed CG samples in comparison with various superalloys reported in the literature [single-phased IN800 (*17*), C276 (*27*), and MP35N (*28*); multiphased A286 (*29*), IN718 (*30*), Nimonic90 (*31*), and Waspaloy (*32*); and single-crystal (SC) CMSX-4 and SRR99 (*33*) superalloys]. Creep data of the SNG-9 at 750°C are also included. Data of CMSX-4/SC are from the minimum uniaxial-tension creep at 750°C at 750 MPa (*2*) and 760°C at 785 MPa (*34*), respectively.

confirming that the GB-networks are stable against the diffusional creep process under gigapascal stress at 700°C.

In a separate experiment, we estimated the GB diffusivity in the SNG-9 and NG-42 samples annealed at 700°C for 100 hours tensioned under a stress of 260 MPa to analyze their creep behaviors comparatively. We determined GB diffusion coefficients in terms of the grain growth kinetic model by using the measured grain size changes. The GB diffusion coefficient we obtained in SNG-9 is about four orders of magnitude smaller than that in NG-42 (*11*), which suggests that atomic diffusion is substantially suppressed by the stable GB networks. This effect echoes previous observations in Cu that atomic diffusion is clearly slowed along the relaxed GBs adjacent to the intersecting twin boundaries (*26*).

Single-phased superalloys usually have low creep stresses—typically below 350 MPa with creep rates below 10^{-6} s^{-1}—and they are higher in the γ′/γ″–strengthened multiphased superalloys at 700°C (400 to ~750 MPa for wrought superalloys and up to ~800 MPa for single-crystal superalloys) (Fig. 4). The creep resistances we observed in the NG-42 and deformed CG samples are obviously lower than those of conventional superalloys (*2*, *17*, *27*–*34*). The creep resistance of the SNG-9 sample is outstanding, with strain rates as low as ~10^{-7} s^{-1} under a gigapascal stress at 700°C and 959 MPa at 750°C. These properties in our single-phased NiCoCr alloy with stable GB networks are superior to the existing wrought single-phased and multiphased superalloys. Our alloy even outperforms that of the most heavily alloyed CMSX-4 (*2*, *34*) and SRR99 (*33*) single-crystal superalloys.

The present finding indicates that the stable GB networks in the nanograined alloy enable an elevated thermal stability, high-temperature strength, as well as high-temperature creep resistance simultaneously. Such a property enhancement differs fundamentally from that of the traditional strategies. We know that over the past decades, various techniques have been developed for introducing high-density GBs in nanostructured materials. Stabilization of GBs through different physical and chemical approaches has been demonstrated in a wide range of metals and alloys (*35*). Therefore, we anticipate that the use of stable GB networks offers a viable paradigm for designing advanced stable alloys with high performance, especially alloys for high-temperature applications.

REFERENCES AND NOTES

1. M. A. Meyers, K. K. Chawla, *Mechanical Behavior of Materials* (Cambridge Univ. Press, 2008).
2. R. C. Reed, *The Superalloys: Fundamentals and Applications* (Cambridge Univ. Press, 2006).
3. K. A. Darling *et al*., *Nature* **537**, 378–381 (2016).
4. R. L. Coble, *J. Appl. Phys.* **34**, 1679–1682 (1963).
5. F. R. N. Nabarro, *Report of a Conference on Strength of Solids* (The Physical Society, 1948).
6. C. Herring, *J. Appl. Phys.* **21**, 437–445 (1950).
7. M. Nganbe, A. Fahim, *J. Mater. Eng. Perform.* **19**, 395–400 (2010).
8. X. Zhou, X. Y. Li, K. Lu, *Science* **360**, 526–530 (2018).
9. X. Zhou, X. Li, K. Lu, *Phys. Rev. Lett.* **122**, 126101 (2019).
10. D. Sorensen, B. Q. Li, W. W. Gerberich, K. A. Mkhoyan, *Acta Mater.* **63**, 63–72 (2014).
11. Materials and methods are available as supplementary materials.
12. K. Oh-ishi, Z. Horita, D. J. Smith, T. G. Langdon, *J. Mater. Res.* **16**, 583–589 (2001).
13. K. Lu, *Nat. Rev. Mater.* **1**, 16019 (2016).
14. W. Xu, B. Zhang, X. Y. Li, K. Lu, *Science* **373**, 683–687 (2021).
15. M. Nganbe, M. Heilmaier, *Int. J. Plast.* **25**, 822–837 (2009).
16. Y. Gu *et al*., *Scr. Mater.* **55**, 815–818 (2006).
17. Special Metals Corporation, "Alloy technical bulletins"; https://www.specialmetals.com/documents/technical-bulletins.
18. G. G. D. Armengol, thesis, University of London (1981).
19. J. P. Hirth, J. Lothe, *Theory of Dislocations* (McGraw-Hill, 1968).
20. J. H. Schneibel, M. Heilmaier, W. Blum, G. Hasemann, T. Shanmugasundaram, *Acta Mater.* **59**, 1300–1308 (2011).
21. T. Xia *et al*., *Mater. Charact.* **145**, 362–370 (2018).
22. I. C. Choi, B. G. Yoo, Y. J. Kim, J. I. Jang, *J. Mater. Res.* **27**, 3–11 (2012).
23. M. A. Meyers, A. Mishra, D. J. Benson, *Prog. Mater. Sci.* **51**, 427–556 (2006).
24. J. Weissmüller, J. Markmann, *Adv. Eng. Mater.* **7**, 202–207 (2005).
25. L. Wang *et al*., *Nat. Commun.* **5**, 4402 (2014).
26. K. C. Chen, W. W. Wu, C. N. Liao, L. J. Chen, K. N. Tu, *Science* **321**, 1066–1069 (2008).
27. X. P. Mao *et al*., *Adv. Mat. Res.* **328-330**, 1143–1148 (2011).
28. R. P. Singh, R. D. Doherty, *Metall. Trans. A Phys. Metall. Mater. Sci.* **23**, 321–334 (1992).
29. H. De Cicco, M. I. Luppo, L. M. Gribaudo, J. Ovejero-García, *Mater. Charact.* **52**, 85–92 (2004).
30. T. W. Ni, J. X. Dong, *Mater. Sci. Eng. A* **700**, 406–415 (2017).
31. G. F. Harrison, W. J. Evans, M. R. Winstone, *Mater. Sci. Technol.* **25**, 249–257 (2009).
32. M. Whittaker, W. Harrison, C. Deen, C. Rae, S. Williams, *Materials* **10**, 61 (2017).
33. G. M. Han, J. J. Yu, Z. Q. Hu, X. F. Sun, *Mater. Charact.* **86**, 177–184 (2013).
34. Y. S. Zhao *et al*., *Mater. Sci. Forum* **944**, 8–12 (2019).
35. C. A. Schuh, K. Lu, *MRS Bull.* **46**, 225–235 (2021).

ACKNOWLEDGMENTS

We thank C. Liang for help with sample preparation. **Funding:** This work was supported by the Ministry of Science and Technology of China (grants 2017YFA0700700 and 2017YFA0204401), the National Natural Science Foundation of China (grant 51901226), and the Chinese Academy of Sciences. **Author contributions:** X.Y.L. and K.L. initiated and supervised the study; Y.G.T. prepared the samples; Q.S.M. performed the indentation creep test; B.B.Z. performed structure characterization experiments; and B.B.Z., X.Y.L., and K.L. analyzed data, discussed the results, and wrote the paper. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are available in the manuscript or the supplementary materials. **License information:** Copyright © 2022 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abq7739
Materials and Methods
Figs. S1 to S6
Tables S1 and S2
References (*36*–*44*)

Submitted 29 April 2022; accepted 11 October 2022
10.1126/science.abq7739

CANCER

TPP1 promoter mutations cooperate with TERT promoter mutations to lengthen telomeres in melanoma

Pattra Chun-on[1,2,3], Angela M. Hinchie[1], Holly C. Beale[4,5], Agustin A. Gil Silva[1], Elizabeth Rush[6], Cindy Sander[6], Carla J. Connelly[7], Brittani K.N. Seynnaeve[6,8], John M. Kirkwood[6], Olena M. Vaske[4,5], Carol W. Greider[4,5,7]*, Jonathan K. Alder[1]*

Overcoming replicative senescence is an essential step during oncogenesis, and the reactivation of *TERT* through promoter mutations is a common mechanism. *TERT* promoter mutations are acquired in about 75% of melanomas but are not sufficient to maintain telomeres, suggesting that additional mutations are required. We identified a cluster of variants in the promoter of *ACD* encoding the shelterin component TPP1. *ACD* promoter variants are present in about 5% of cutaneous melanoma and co-occur with *TERT* promoter mutations. The two most common somatic variants create or modify binding sites for E-twenty-six (ETS) transcription factors, similar to mutations in the *TERT* promoter. The variants increase the expression of *TPP1* and function together with TERT to synergistically lengthen telomeres. Our findings suggest that *TPP1* promoter variants collaborate with *TERT* activation to enhance telomere maintenance and immortalization in melanoma.

Escaping replicative senescence is an essential step of oncogenesis (*1*, *2*). Telomere shortening limits the proliferative potential of cells, and several mechanisms have been identified that permit tumor cells to extend telomeres and increase their replicative capacity (*3*–*8*). Somatic mutations in the *TERT* promoter are the most common identifiable mechanism in melanoma and are found in ~75% of cases (*5*, *9*). TERT is the catalytic component of telomerase, the enzyme responsible for de novo telomere synthesis and maintenance of telomeres. *TERT* promoter mutations are not sufficient to immortalize some cell types, and telomeres continue to shorten in nevi during the transition to melanoma despite acquisition of *TERT* promoter mutations (*10*). Therefore, although additional mutations are likely required to enable telomere elongation and immortalization, the genomic changes that potentially synergize with *TERT* promoter mutations to achieve sustained telomere maintenance remain unknown (*10*).

[1]Dorothy P. and Richard P. Simmons Center for Interstitial Lung Disease, Division of Pulmonary, Allergy, and Critical Care Medicine, Pittsburgh, PA, USA. [2]Environmental and Occupational Health Department, School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA. [3]Faculty of Medicine and Public Health, Princess Srisavangavadhana College of Medicine, Chulabhorn Royal Academy, Bangkok, Thailand. [4]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, CA, USA. [5]Department of Molecular, Cell and Developmental Biology, University of California, Santa Cruz, CA, USA. [6]University of Pittsburgh Medical Center, Hillman Cancer Institute, Pittsburgh, PA, USA. [7]Department of Molecular Biology and Genetics, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [8]Department of Pediatrics, Division of Pediatric Hematology/Oncology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA.
*Corresponding author. Email: jalder@pitt.edu (J.K.A.) and cgreider@ucsc.edu (C.W.G.)

We sought to identify new mechanisms of telomere maintenance in cancer cells by analyzing somatic mutations that occur in telomere-related genes. We focused our analysis on melanoma because individuals with germline variants in the *TERT* promoter predominantly develop this cancer (*5*). We examined somatic variants from telomere-related genes in melanoma from the International Cancer Genome Consortium (ICGC) (*11*). Mutations in TPP1 and POT1, components of the six-protein shelterin complex that coats telomeres, have been reported in familial melanoma (*12*–*14*), and we found numerous somatic variants among 749 melanoma samples analyzed (fig. S1). We noted a cluster of recurrent somatic variants in a conserved region of *ACD*, the gene encoding TPP1 (which we refer to as the *TPP1* gene hereafter for clarity), that co-localized with histone marks typically associated with promoters (Fig. 1A). *TPP1* has been reported to have two isoforms, TPP1-long (L) and TPP1-short (S), which differ by 86 amino acids in the N terminus (*15*, *16*). The cluster of variants was positioned such that they would be coding variants in TPP1-L and promoter variants in TPP1-S. To determine which isoform of TPP1 was expressed in melanoma, we examined RNA-sequencing (RNA-seq) data from 12 melanoma cell lines and 61 microdissected nevi and melanoma samples (GSE153592 and GSE112509) (*17*, *18*), and found that TPP1-S was the only isoform expressed (Fig. 1A and fig. S2). To further validate this finding, we cloned the entire genomic region [3.5 kb, including 895 base pairs (bp) upstream of the TPP1-S translational start site and 637 bp upstream of the TPP1-L translation start site and all exons and introns of both isoforms] of *TPP1* into a plasmid without an additional promoter and incorporated a C-terminal FLAG-tag. Immunofluorescent staining confirmed that the C-terminally tagged TPP1 co-localized with TRF2 at telomeres (Fig. 1B). We generated cDNA expression constructs for TPP1-L, TPP1-S, and TPP1-L(M87A), which is incapable of expressing TPP1-S, as controls (*15*). Western blot of cells expressing the entire genomic region with (TPP1pro −108C>T and TPP1pro −75C>T) and without (TPP1pro WT) the two most common variants showed that only TPP1-S was expressed in HEK293 and the melanocytic cell lines LOX and MEL624 (Fig. 1C). Together, these findings support the idea that the cluster of variants that we identified are localized to the promoter of TPP1-S. For clarity, we will refer to TPP1-S as TPP1 hereafter.

Previous studies have investigated the role of noncoding mutations in cancers including melanoma (*19*–*21*). Because the annotations of 5′ portion of *TPP1* have changed from coding to noncoding in recent years, (fig. S2), the region that we identified would not have been included as a promoter region in earlier studies. We tested whether the proximal 200 bp upstream of the *TPP1* translational start site was enriched for somatic variants by examining whole-genome sequencing data from 305 patient-derived melanoma samples available from the ICGC, and found that the *TPP1* proximal promoter was significantly enriched relative to 59,727 annotated promoters (false discovery rate–corrected P value = 6.57×10^{-14}).

We next sought to determine the functional consequences of the somatic variants that we identified in the promoter of *TPP1*. The two most common variants were C>T transitions located 75 and 108 bp upstream of the translational start site. The −108 variant created the core binding sequence TTCC for the E-twenty-six (ETS) family of transcription factors. The −75 variant was adjacent to an existing ETS site in the context of a sequence that is enriched for mutations in melanoma (*21*), created a new TFIID binding site, and co-localized with the annotated transcriptional start site for the TPP1 mRNA (Fig. 2 and fig. S3). The identification of two recurrent promoter variants that created or modified ETS transcription factor binding sites bore marked similarity to putative activating mutations in the *TERT* promoter (*5*, *6*), although the precise sequences created by the variants were distinct. We found that *TPP1* was up-regulated in several large databases of cancer gene expression, including those with a high frequency of *TERT* promoter mutations (fig. S4); however, there were insufficient data for the analysis of melanoma with *TPP1* promoter mutations specifically. To further characterize the *TPP1* promoter variant, we generated luciferase reporters of progressively

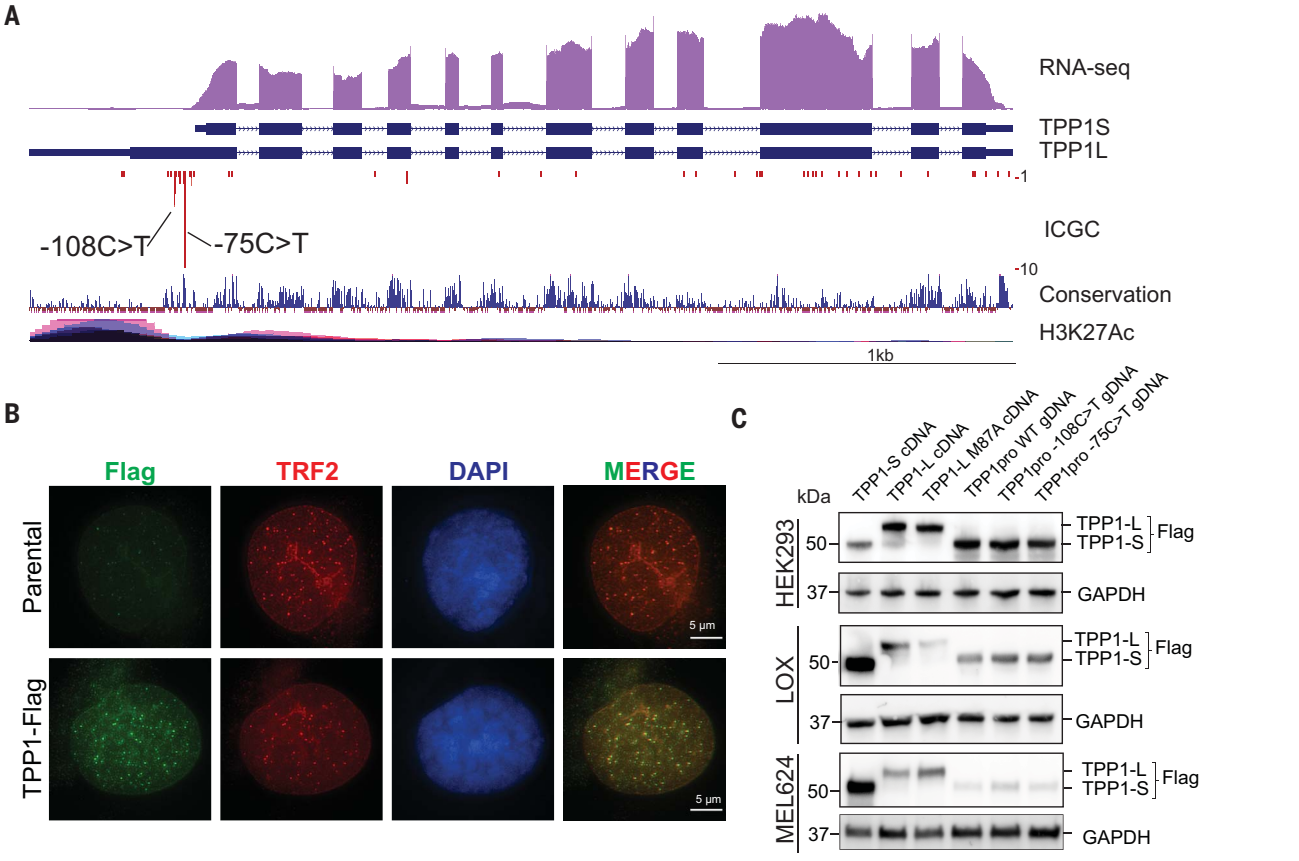


Fig. 1. Identification of a cluster of somatic promoter variants in *TPP1*. (**A**) Genomic locus of the *ACD* gene from UCSC Genome Browser data. Dark blue rectangles indicate the exons for TPP1-S and TPP1-L. The red bars below the gene track show the locations of the somatic variants identified in the ICGC database, with taller bars corresponding to the number of melanomas found with a specific variant. RNA-seq data (GSE153592) are shown above the gene track in purple, along with vertebrate conservation and H3K27 acetylation marks from multiple cell lines, indicating the location of likely regulatory regions. (**B**) HeLa cell lines stably expressing a C-terminally FLAG-tagged TPP1 were stained for the shelterin component TRF2 and the FLAG epitope. Co-localization of TPP1 with TRF2 suggests that the C-terminal FLAG-tag does not disrupt localization of TPP1 to the telomere. (**C**) Western blot of HEK293, LOX, and MEL624 cells transfected with plasmids encoding the cDNAs for TPP1-S and TPP1-L and for TPP1-L-M87A, which is incapable of expressing TPP1-S, together with plasmids expressing the entire genomic locus of TPP1 with and without the most common promoter variants.

smaller fragments of the TPP1 proximal promoter, and found that a 285-bp fragment was sufficient for full basal transcriptional activity (fig. S3E). Introduction of the −75C>T or −108C>T promoter variants had little effect on luciferase expression in HEK293 cells; however, there was a small but significant increase in luciferase expression in the two melanoma cells lines, LOX and MEL624 (Fig. 2A), suggesting that melanoma-specific ETS transcription factors are required for increased activity. RNA-seq data from 426 melanoma samples (*11*) showed that of 27 ETS family members, ETV5, ETS1, and ETV4 were the most abundantly expressed (fig. S5). We confirmed this finding using quantitative polymerase chain reaction (qPCR) and Western blotting on non-melanoma lines HeLa, BJ fibroblasts, and HEK293 and several melanoma cell lines and short-term primary cultures (Fig. 2, B and C). Finally, we overexpressed ETV5, ETV4, or ETS1 in HEK293 cells and found that all three robustly increased the activity of the TPP1 promoter only when the promoter variants were present (Fig. 2D). These data suggest that the *TPP1* promoter variants are activated by ETS transcription factors that are abundantly expressed in melanomas.

To determine the cellular consequences of increased TPP1 expression, we generated stable cell lines that overexpressed TPP1 in telomerase-positive HeLa cells. For completeness, we also examined TPP1-L. We investigated whether cells stably expressing C-terminally tagged TPP1-S and TPP1-L altered telomere length in telomerase-expressing HeLa cells. Consistent with previous reports (*15*, *22*), stable overexpression of TPP1-S led to considerable telomere lengthening, whereas overexpression of TPP1-L caused telomere shortening (Fig. 3A). These data confirm that increased expression of TPP1 can lead to telomere lengthening in telomerase-expressing cells.

We next investigated whether the overexpression of TPP1 could extend the proliferative capacity of cells that express limiting amounts of telomerase. We expressed TPP1, TERT, or both in primary BJ fibroblasts and monitored their proliferative capacity for 90 days (Fig. 3, B and C). Control untransduced fibroblasts and fibroblasts transduced with TPP1 alone entered replicative senescence after ~40 days. By contrast, cells transduced with TERT and TERT + TPP1 bypassed senescence and were immortalized (Fig. 3B). These findings are consistent with previous reports demonstrating that TERT overexpression is sufficient to immortalize primary fibroblasts (*23*). We next examined telomere length in control and TPP1-transduced fibroblasts, and found that telomeres were very heterogeneous, with a median length of ~6 kb after 15 passages. Introduction of TERT caused telomere lengthening, consistent with previous reports (*23*), and the coexpression of TERT and TPP1 together caused a synergistic effect resulting in significant telomere elongation (Fig. 3D). These findings

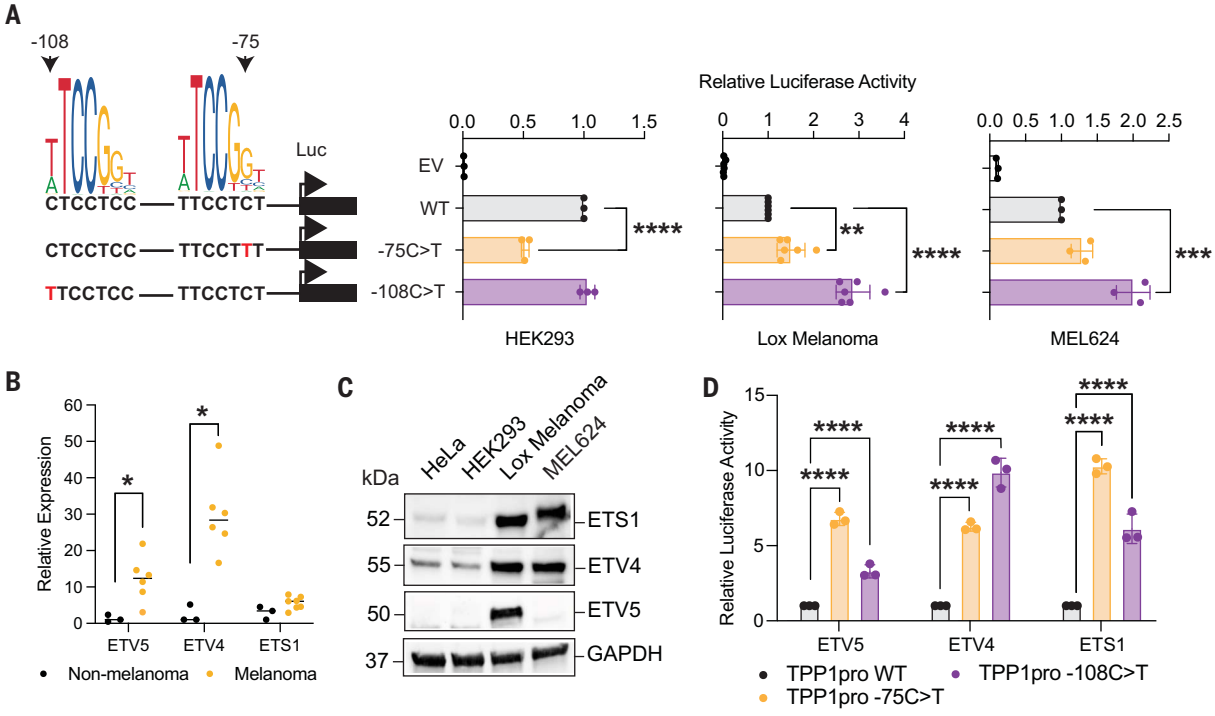


Fig. 2. ETS transcription factors activate the variant TPP1 promoter. (**A**) Luciferase assays were performed with a 285-bp fragment of the TPP1 proximal promoter in melanoma and non-melanoma cell lines. The TPP1 promoter variants had little effect on the transcriptional activity in non-melanoma cells lines (HEK293) but increased reporter activity in two melanoma-derived lines. (**B**) qPCR examining the levels of three ETS transcription factor family members in non-melanoma (HeLa, HEK293, and BJ fibroblast; $n = 3$) and melanoma cell lines (MEL624 and LOX) and short-term primary cultures ($n = 6$ to 7). Medians are shown and groups were compared using the Mann-Whitney test. (**C**) Western blot showing high expression of ETS transcription factors in LOX melanoma and MEL624 lines. (**D**) Luciferase assays comparing activity of the TPP1 promoter reporter in the presence of three transfected ETS transcription factors in HEK293 cells. Cells were co-transfected with a pGL4 reporter and pCDNA3.1 expression plasmid with one of the three ETS transcription factors. Mean and SD are shown from at least three independent experiments in (A) and (D), and groups were compared with a one-way ANOVA followed by Tukey's multiple-comparisons test for pairwise comparisons. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, and ****$P < 0.0001$.

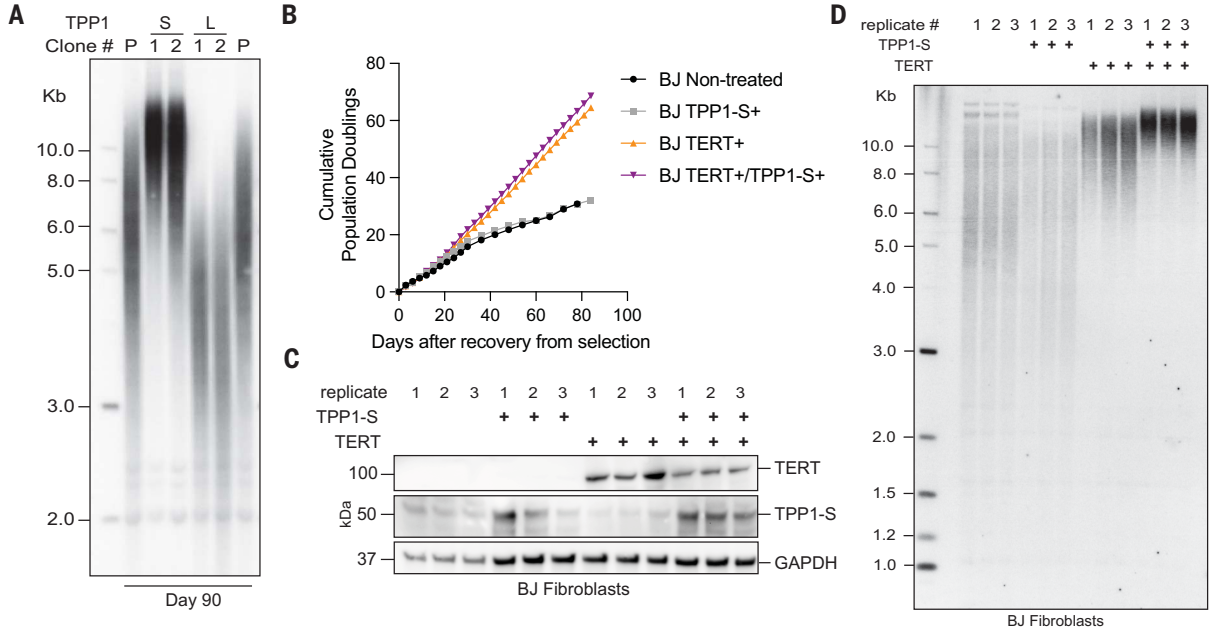


Fig. 3. TPP1-S overexpression causes telomere lengthening and is synergistic with TERT overexpression. (**A**) Southern blot of telomeres in HeLa cell lines that stably expressed TPP1-S or TPP1-L for 90 days. Two independent clones of each are shown. "P" indicates the parental cell line used to establish each of the modified clones. (**B**) Growth curves of cumulative population doublings of BJ fibroblasts expressing TPP1-S or TERT (average of three independent transductions for each group). (**C**) Western blot showing expression of each of the transgenes in cells collected from (B). (**D**) Southern blot of telomere lengths of BJ fibroblasts in (B) 15 passages after transduction showing synergistic telomere lengthening in cells exogenously overexpressing TPP1 and TERT.

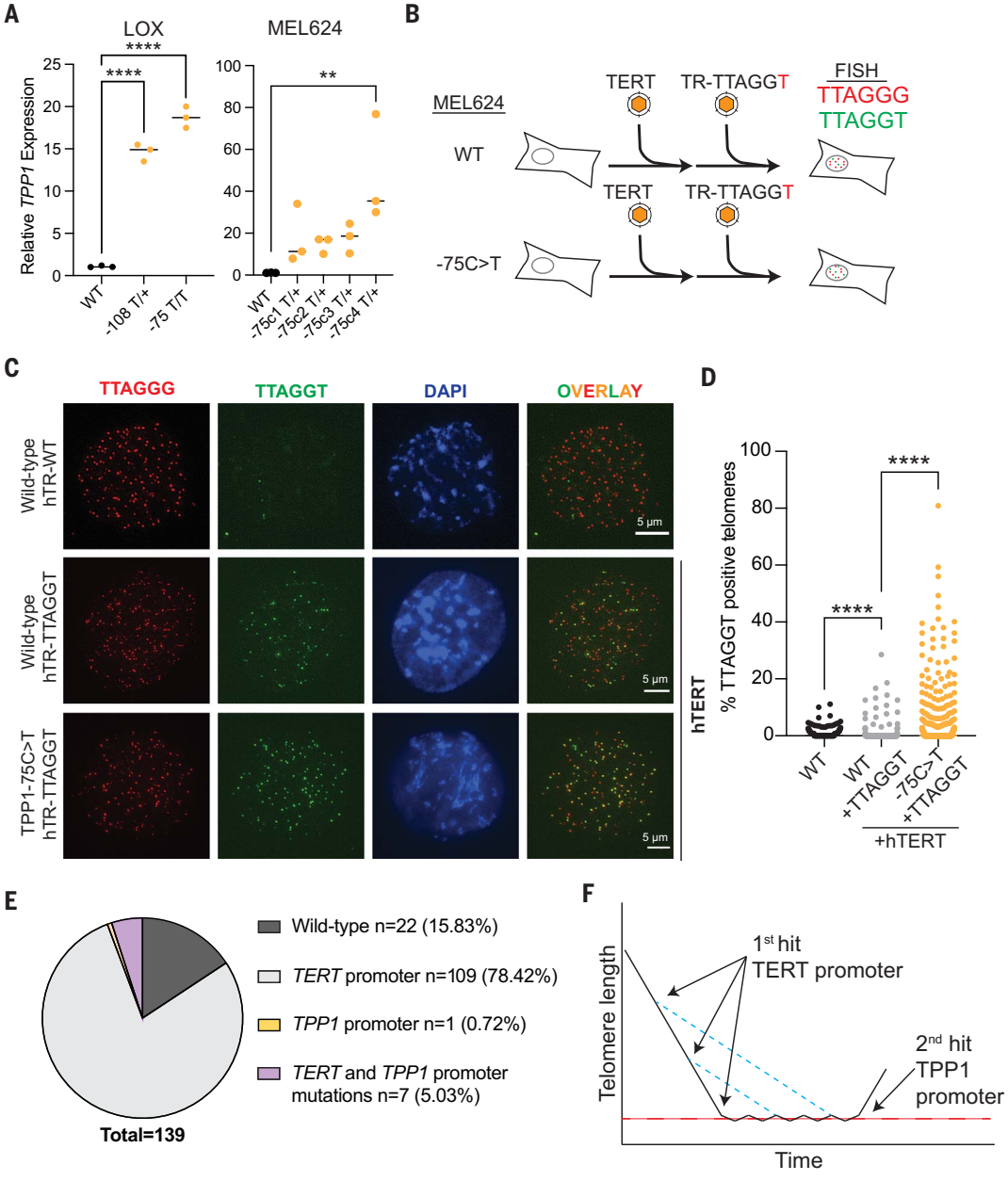


Fig. 4. TPP1 promoter mutations increase the expression of the endogenous transcript and co-occur with *TERT* promoter mutations. (**A**) qPCR of *TPP1* expression after the introduction of promoter mutations in LOX and MEL624 cells. Labels below the graph indicate the presumed zygosity based on sequencing. The median is shown from three independent measurements from each clone, and groups were compared using one-way ANOVA followed by Dunnett's multiple-comparisons test. (**B**) Schematic of the experimental approach to measuring telomerase activity in genetically modified cells. Cells were transduced with a *TERT*-expressing lentivirus to increase the rate of variant telomere incorporation. After the introduction of the mutant telomerase RNA (encoding TTAGGT), cells were passaged and the canonical and variant telomeres were quantitated. (**C**) FISH for the WT (TTAGGG; red) and variant (TTAGGT; green) in parental or genome-edited MEL624 cells. Images were taken 7 days after transduction with lentiviruses. (**D**) Quantitation of the fraction of telomeres with both TTAGGG and TTAGGT signals from a single clone. Groups were compared using ANOVA with Dunnett's correction for multiple comparisons. $^{**}P < 0.01$ and $^{****}P < 0.0001$. (**E**) Proportion of cutaneous melanomas that had *TERT*, *TPP1*, or *TERT* + *TPP1* variants (*25*). (**F**) Model of telomere length dynamics in melanoma progression. *TERT* promoter variants likely occur early and slow telomere attrition but are not sufficient to prevent telomere shortening (blue dashed lines in model). Telomere shortening continues until cells enter crisis (red dashed line). Additional mutations, such as the *TPP1* promoter, are likely required to fully maintain telomeres and escape crisis (second hit).

indicate that TERT and TPP1 overexpression is synergistic and lengthens telomeres more than TERT overexpression alone.

To determine whether the TPP1 promoter variants were sufficient to increase telomere addition, we introduced the two most common variants, –75C>T and –108C>T, into MEL624 and LOX cells using CRISPR/Cas9 (see the supplementary materials and methods) (fig. S6). We obtained and sequence verified six clones from MEL624 and LOX cells. We then examined the expression of *TPP1* after modifying the endogenous promoter and found that introduction of either the –75C>T or –108C>T variant significantly increased the expression of *TPP1* (Fig. 4A). The greater increase in expression levels of *TPP1* from the modified endogenous promoter (Fig. 4C) compared with the luciferase assay (Fig. 2A) suggests that additional factors may contribute to *TPP1* expression in melanoma. Telomeres are extremely long in MEL624 and LOX cells (>20 kb), and it is not possible to detect changes in length using Southern blot analysis. We therefore used fluorescence in situ hybridization (FISH) to detect a modified telomere sequence as a surrogate for in vivo telomerase activity, as previously described (Fig. 4B) (*24*). We used a telomerase RNA encoding the variant telomere repeat TTAGGT, which can be incorporated into telomeres and localized with a peptide nucleic acid fluorescent probe. Wild-type or genome-edited MEL624 or LOX cell lines with the most common promoter variants were co-transduced with lentiviruses that express the variant telomerase RNA and hTERT. Using FISH to determine the percentage of telomeres with variant repeats, we found that cells with a modified *TPP1* promoter incorporated significantly more TTAGGT variant repeat sequences on telomeres (Fig. 4, C and D, and fig. S7). These findings suggest that

TPP1 promoter mutations synergize with hTERT to increase telomere repeat addition in melanoma cells.

We next investigated the co-occurrence of somatic *TERT* and *TPP1* promoter mutations in cancer. *TPP1* promoter variants are found primarily in cancers of the skin, but have also been reported in several different cancer types (fig. S8). The reason for the disproportionate number of variants in melanoma is unknown but may be related to the high mutation rate and reliance on telomerase activation in this cancer. In a dataset of deeply sequenced cutaneous melanomas (*25*), 139 samples were evaluated for the presence of the *TERT* and *TPP1* promoter variants. A large fraction (83%) carried a somatic variant in the *TERT* promoter, as previously reported (*5*, *9*), and eight samples carried a TPP1 promoter variant (~6%). In all cases except one, the *TERT* and *TPP1* promoter variants were found together in the same tumor (Fig. 4E). However, because of limitations in current whole-genome sequencing datasets, additional studies using targeted resequencing will be required to determine the frequency of *TPP1* and *TERT* promoter mutations in cancers other than melanoma. We found that TPP1 up-regulation in the absence of telomerase expression is unlikely to influence telomere length or cellular immortality. Therefore, selection for *TPP1* promoter variants is most likely to occur after the activation of telomerase (Fig. 4E). Our data indicate that TPP1 is one of the missing factors that collaborate with *TERT* promoter mutations to facilitate cellular immortalization in melanoma. The identification of new pathways that contribute to telomere lengthening and cellular immortalization may have important prognostic value and may also inform the development of possible treatments for patients with cancer and those with diseases of telomere shortening. Our findings also support the idea that multiple noncoding mutations can cooperate to enable cellular immortalization and highlight the importance of understanding the contribution of noncoding variants to the development of cancer.

REFERENCES AND NOTES

1. N. W. Kim *et al.*, *Science* **266**, 2011–2015 (1994).
2. D. Hanahan, R. A. Weinberg, *Cell* **144**, 646–674 (2011).
3. C. B. Harley, A. B. Futcher, C. W. Greider, *Nature* **345**, 458–460 (1990).
4. K. J. Wu *et al.*, *Nat. Genet.* **21**, 220–224 (1999).
5. S. Horn *et al.*, *Science* **339**, 959–961 (2013).
6. F. W. Huang *et al.*, *Oncogenesis* **4**, e176 (2015).
7. F. P. Barthel *et al.*, *Nat. Genet.* **49**, 349–357 (2017).
8. M. Peifer *et al.*, *Nature* **526**, 700–704 (2015).
9. F. W. Huang *et al.*, *Science* **339**, 957–959 (2013).
10. K. Chiba *et al.*, *Science* **357**, 1416–1420 (2017).
11. J. Zhang *et al.*, *Nat. Biotechnol.* **37**, 367–369 (2019).
12. C. D. Robles-Espinoza *et al.*, *Nat. Genet.* **46**, 478–481 (2014).
13. L. G. Aoude *et al.*, *J. Natl. Cancer Inst.* **107**, dju408 (2014).
14. J. Shi *et al.*, *Nat. Genet.* **46**, 482–486 (2014).
15. S. Grill *et al.*, *Cell Rep.* **27**, 3511–3521.e7 (2019).
16. J. M. Boyle *et al.*, *Mol. Biol. Cell* **31**, 2583–2596 (2020).
17. J. Motwani *et al.*, *Epigenomics* **13**, 577–598 (2021).
18. M. Kunz *et al.*, *Oncogene* **37**, 6136–6151 (2018).
19. E. Rheinbay *et al.*, *Nature* **578**, 102–111 (2020).
20. N. Weinhold, A. Jacobsen, N. Schultz, C. Sander, W. Lee, *Nat. Genet.* **46**, 1160–1165 (2014).
21. N. J. Fredriksson *et al.*, *PLOS Genet.* **13**, e1006773 (2017).
22. J. Nandakumar *et al.*, *Nature* **492**, 285–289 (2012).
23. A. G. Bodnar *et al.*, *Science* **279**, 349–352 (1998).
24. M. E. Diolaiti, B. A. Cimini, R. Kageyama, F. A. Charles, B. A. Stohr, *Nucleic Acids Res.* **41**, e176 (2013).
25. N. K. Hayward *et al.*, *Nature* **545**, 175–180 (2017).

ACKNOWLEDGMENTS

We thank all patients who participated in the various studies reported here; the Melanoma Center Biospecimen Repository (UPCI 96-99) for providing samples; C. Wongchokprasitti and M. Diekans for assistance in bioinformatic analysis and data presentation; and P. Opresko and T. Finkel for thoughtful feedback on this manuscript. **Funding:** This work was supported by the National Institutes of Health (grant R35CA209974 to C.W.G. and J.K.A. and grant R01HL135062 to J.K.A.). P.C. received funding from the HRH Princess Chulabhorn College of Medical Science to support graduate studies in Environmental and Occupational Health Science, School of Public Health, University of Pittsburgh. **Author contributions:** Conceptualization: C.W.G., J.K.A.; Data curation: J.K.A., P.C., H.B.; Formal analysis: J.K.A., P.C., H.B., O.V.; Funding acquisition: C.W.G., J.K.A.; Investigation: P.C., B.S., J.M.K., C.W.G., J.K.A., H.C.B., O.M.B.; Methodology: P.C., A.M.H., A.A.G.S., E.R., C.S., C.J.C., H.C.B., O.M.B.; Visualization: P.C., A.M.H., H.C.B.; Resources: E.R., C.S., B.S., J.M.K.; Writing – original draft: P.C., C.W.G., J.K.A.; Writing – review and editing: all authors. **Competing interests:** J.M.K. reports advisory/consultancy roles with Amgen Inc., Ankyra Therapeutics, Applied Clinical Intelligence LLC, Axio Research LLC, Becker Pharmaceutical Consulting, Bristol Myers Squibb, Cancer Network, Checkmate Pharmaceuticals, DermTech, Fenix Group International, Harbour BioMed, Immunocore LLC, iOnctura, Istari Oncology, Magnolia Innovations LLC, Merck, Natera Inc., Novartis Pharmaceuticals, OncoCyte Corporation, OncoSec Medical Inc., PathAI Inc., Pfizer Inc., Replimune, Scopus BioPharma, SR One Capital Management, Takeda Development Center Americas Inc., and Takeda Pharmaceutical Company Limited. J.M.K. also reports research grants/funding to his institution from Amgen Inc., Bristol Myers Squibb, Checkmate Pharmaceuticals, Harbour BioMed, Immvira Pharma Co., Immunocore LLC, Iovance Biotherapeutics, Novartis Pharmaceuticals, Takeda, and Verastem Inc. The remaining authors declare no competing interests. **Data and materials availability:** All data and materials used in the analysis are available to any researcher for purposes of reproducing or extending our analysis and are available in the main text or the supplementary materials. Please contact the corresponding authors to request any materials. **License information:** Copyright © 2022 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abq0607
Materials and Methods
Figs. S1 to S8
References (*26*–*39*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

Submitted 24 March 2022; accepted 11 October 2022
10.1126/science.abq0607

TECHNICAL COMMENT

ENVIRONMENTAL TOXINS

Comment on "Models predict planned phosphorus load reduction will make Lake Erie more toxic"

Jef Huisman[1]*, Elke Dittmann[2], Jutta Fastner[3], J. Merijn Schuurmans[1], J. Thad Scott[4], Dedmer B. Van de Waal[1,5], Petra M. Visser[1], Martin Welker[6], Ingrid Chorus[3]

Hellweger *et al.* (Reports, 27 May 2022, pp. 1001) predict that phosphorus limitation will increase concentrations of cyanobacterial toxins in lakes. However, several molecular, physiological, and ecological mechanisms assumed in their models are poorly supported or contradicted by other studies. We conclude that their take-home message that phosphorus load reduction will make Lake Erie more toxic is seriously flawed.

Toxic cyanobacterial blooms cause major water quality problems across the globe. Hence, there is a need for models that can predict cyanotoxin concentrations in surface waters. Hellweger *et al.* (*1*) developed an agent-based model that can provide valuable insights in the molecular and physiological mechanisms affecting the toxicity of cyanobacterial blooms. In essence, the proposed mechanisms in Hellweger *et al.* (*1*) are: (i) High nitrogen (N) but low phosphorus (P) loads will stimulate the production of the N-rich toxin microcystin (MC). (ii) One of the key functions of MC is protection against oxidative stress. (iii) Therefore, MC-producing strains will be more resistant to natural H_2O_2 concentrations than non-MC-producing strains, especially under high N but low P conditions. (iv) Thus, reducing P loads without diminishing N loads will select for MC-producing strains, that will make Lake Erie "more toxic".

Hellweger *et al.* suggest that these mechanisms have contributed to the observed resurgence of toxic cyanobacteria after P load reduction in Lake Erie and many other lakes, and they advocate a dual N and P management strategy. Although we agree with Hellweger *et al.* that high N concentrations can be an important driver of cyanobacterial growth and cyanotoxin production (*2*,*3*), several of their model assumptions and predictions are poorly supported or contradicted by existing literature.

First, contrary to the claim in their title, the model of Hellweger *et al.* does not make predictions about toxicity. MC comprises a large class of cyanobacterial toxins, consisting of hundreds of MC congeners that vary widely in toxicity. Their model considers only the total MC concentration, but ignores changes in MC composition and therefore cannot make predictions about the toxicity of blooms. This is not merely a semantic issue, because excess N may shift the MC composition to the more N-rich variant MC-RR (*2*), which is one of the least toxic MC congeners.

Second, contrary to their Fig. 2, cyanobacteria produce only low amounts of H_2O_2 by photosynthesis. Cyanobacteria lack the Mehler reaction, which is responsible for most H_2O_2 production during high light stress in photosynthetic eukaryotes. Instead, cyanobacteria use a "Mehler-like" reaction with flavodiiron proteins to transfer their excess photosynthetic electrons to O_2, which produces water without H_2O_2 formation (*4*).

Third, Hellweger *et al.* assume that the natural H_2O_2 concentrations in Lake Erie (0.1 to 0.5 μmol/L) may cause oxidative stress and induce MC binding to proteins. However, the study (*5*) cited by Hellweger *et al.* used a much higher H_2O_2 concentration of 10 μmol/L to induce MC binding to proteins. Dziallas and Grossart (*6*) reported significant reduction of the cellular chlorophyll-*a* content at environmentally relevant H_2O_2 concentrations of 0.025 to 0.1 μmol/L, but this result appears to deviate from other studies. Most controlled laboratory studies show that H_2O_2 only starts to affect the photosynthetic yield and growth of *Microcystis* strains at H_2O_2 concentrations that are one or more orders of magnitude higher [5 to 60 μmol/L, depending on the conditions; (e.g., *7*,*8*)]. This is in agreement with lake treatments that require H_2O_2 concentrations of 60 to 300 μmol/L to effectively suppress cyanobacterial blooms (*8*,*9*). We therefore question whether the natural H_2O_2 concentrations in Lake Erie are high enough to induce MC-binding to proteins and to shift the competitive balance between toxic and nontoxic *Microcystis* strains.

Fourth, the function of MC in cyanobacterial cells has remained elusive for decades. For example, MC has been implicated in grazing defense, allelopathic interactions, iron scavenging, protection against oxidative stress, carbon-nitrogen metabolism, and cell signaling (*10*). Hellweger *et al.* adhere to the hypothesis (*5*) that MC binding protects proteins such as RuBisCO against oxidative stress, which would provide a selective advantage to MC-producing strains when exposed to H_2O_2. Binding of MC to proteins has indeed been unequivocally demonstrated (*5*). However, whether MC-producing cells are better protected against H_2O_2 is less clear and contradicted by other experiments (*11*). Recent work from the research group that originally proposed the "protection against oxidative stress hypothesis" indicates that MC binding to RuBisCO probably serves a very different function. Binding of MC appears to play a key role in the assembly and cellular localization of RuBisCO, which enables rapid acclimation of cells to CO_2-limiting conditions (*12*).

Fifth, which factors govern the competition between toxic and non-toxic strains? Laboratory selection experiments have shown that the MC-producing wildtype has a strong selective advantage compared to the MC-deletion mutant under CO_2-limited conditions, but not under CO_2-replete conditions (13). This reinforces the recent idea (12) that MC binding to RuBisCO probably plays a role in CO_2 fixation. Other selection experiments investigated the role of N and P limitation. The results showed that the MC-producing wildtype won under N limitation, while the non-toxic mutant dominated under P-limited conditions (14), which is exactly opposite to the predictions of Hellweger *et al.*

Sixth, the models of Hellweger *et al.* consider only toxic and nontoxic *Microcystis* strains. However, the main aim of nutrient load reduction programs is not to shift the competitive balance between *Microcystis* strains, but to suppress the entire cyanobacterial bloom and shift the lake to a completely different phytoplankton community. Shifts from bloom-forming cyanobacteria to (nontoxic) eukaryotic phytoplankton are often observed in response to declines in nutrient availability. Models ignoring this important ecological mechanism are therefore unlikely to make reliable predictions of how nutrient load reductions will affect cyanotoxin concentrations in surface waters.

P load reductions have successfully controlled cyanobacterial blooms in a wide variety of lakes (15), but in some lakes achieving sufficiently low P concentrations proves challenging. In these cases, bloom control by dual N and P reductions seems promising and further experience with such dual approaches is desirable. However, there are major issues with several molecular and physiological mechanisms assumed in Hellweger *et al.* and their models omit common ecological responses to nutrient load reduction

[1]Department of Freshwater and Marine Ecology, Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, P.O. Box 94240, 1090 GE Amsterdam, Netherlands. [2]Department of Microbiology, Institute of Biochemistry and Biology, University of Potsdam, Karl-Liebknecht-Str. 24/25, 14476 Potsdam-Golm, Germany. [3]Department of Drinking Water and Swimming Pool Hygiene, German Environment Agency, Berlin, Germany. [4]Department of Biology, Center for Reservoir and Aquatic Systems Research, Baylor University, One Bear Place #97388, Waco, Texas, USA. [5]Department of Aquatic Ecology, Netherlands Institute of Ecology (NIOO-KNAW), Wageningen, Netherlands. [6]Independent Consultant, Berlin.
*Corresponding author. Email: j.huisman@uva.nl

such as shifts in phytoplankton species composition. Hence, there is insufficient support for their provocative claim that P load reduction alone will make Lake Erie and other lakes more toxic.

REFERENCES AND NOTES

1. F. L. Hellweger *et al*., *Science* **376**, 1001–1005 (2022).
2. D. B. Van de Waal *et al*., *Ecol. Lett.* **12**, 1326–1335 (2009).
3. C. J. Gobler *et al*., *Harmful Algae* **54**, 87–97 (2016).
4. Y. Helman *et al*., *Curr. Biol.* **13**, 230–235 (2003).
5. Y. Zilliges *et al*., *PLOS ONE* **6**, e17615 (2011).
6. C. Dziallas, H. P. Grossart, *PLOS ONE* **6**, e25569 (2011).
7. M. Drábková, W. Admiraal, B. Marsálek, *Environ. Sci. Technol.* **41**, 309–314 (2007).
8. E. F. J. Weenink *et al*., *Environ. Microbiol.* **23**, 2404–2419 (2021).
9. H. C. P. Matthijs *et al*., *Water Res.* **46**, 1460–1472 (2012).
10. A. Holland, S. Kinnear, *Mar. Drugs* **11**, 2239–2258 (2013).
11. J. M. Schuurmans *et al*., *Harmful Algae* **78**, 47–55 (2018).
12. T. Barchewitz *et al*., *Environ. Microbiol.* **21**, 4836–4851 (2019).
13. D. B. Van de Waal *et al*., *ISME J.* **5**, 1438–1450 (2011).
14. S. Suominen, V. S. Brauer, A. Rantala-Ylinen, K. Sivonen, T. Hiltunen, *Aquat. Ecol.* **51**, 117–130 (2017).
15. J. Fastner *et al*., *Aquat. Ecol.* **50**, 367–383 (2016).

ACKNOWLEDGMENTS

Author contributions: Conceptualization: all authors; Writing of original draft: J.H., M.W., and I.C.; Review and editing of manuscript: all authors. **Competing interests:** The authors declare that they have no competing interests.

License information: Copyright © 2022 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.sciencemag.org/about/science-licenses-journal-article-reuse

Submitted 18 July 2022; accepted 12 October 2022
10.1126/science.add9959

TECHNICAL RESPONSE

ENVIRONMENTAL TOXINS

Response to Comment on "Models predict planned phosphorus load reduction will make Lake Erie more toxic"

Ferdi L. Hellweger[1]*†, Charlotte Schampera[1]†, Robbie M. Martin[2], Falk Eigemann[1], Derek J. Smith[3], Gregory J. Dick[3,4], Steven W. Wilhelm[2]*

Huisman *et al*. claim that our model is poorly supported or contradicted by other studies and the predictions are "seriously flawed." We show their criticism is based on an incomplete selection of evidence, misinterpretation of data, or does not actually refute the model. Like all ecosystem models, our model has simplifications and uncertainties, but it is better than existing approaches hat ignore biology and do not predict toxin concentration.

Huisman *et al*. provide a point-by-point criticism of our recent paper (*1*), to which we respond in turn. First, the model predicts total MC concentration and we equate that to toxicity, which is necessitated by limited availability of congener-specific knowledge on synthesis, function, potency, and it is common practice, e.g., WHO, EPA, ELISA, also by Huisman *et al*. (*2*). We agree that a change in N-limitation will likely affect MC congener composition and toxicity. However, the underlying mechanisms remain unclear and the evidence is not as consistent as Huisman *et al*. suggest. In the study they cite (*3*), the more toxic MC-YR also increased, and some studies found no change of composition with changing N (*4*).

Second, in the model, H_2O_2 is produced photosynthetically by *Microcystis* and other sources, including respiration and extracellular. Yes, there is increasing evidence that cyanobacteria do not produce H_2O_2 via the Mehler reaction, but photosynthetic production has been observed (*5*). In Lake Erie, H_2O_2 peaks prior to or coincident with *Microcystis* blooms (*6*, *7*), consistent with photosynthetic production. However, our more recent work also suggests most biological H_2O_2 production in Lake Erie is by heterotrophic bacteria (*7*). It would be useful to extend the model to more explicitly resolve the various H_2O_2 sources and sinks. In the meantime, the simplified representation is a reasonable approximation, and if it is *Microcystis* or associated bacteria is not critical since H_2O_2 readily diffuses across cell membranes.

Third, Huisman *et al*. question whether oxidative stress occurs at environmental H_2O_2 concentrations and cite two studies. The first study (*8*) does not support their assertion. The lowest concentration evaluated was 15 μmol/L at which effects were observed, which does not rule out effects at untested lower concentrations, which were also observed in another study by the same group (*9*). The second study (*10*) found no effect at 22 μmol/L, but the H_2O_2 was degraded within a few hours and not replenished as it was in a study that did observe an effect (*11*) and as it would be in the environment. Our more recent work also shows effects at natural H_2O_2 concentrations, though it also suggests more complex strain-level diversity of H_2O_2 sensitivity (*6*). Note that the effects in the model are sublethal and subtle, corresponding to ~20% growth rate differences. Huisman *et al*. also question whether MC binds to proteins at ambient conditions, although their own work (*12*) showed this.

Fourth, Huisman *et al*. point to a study (*13*) that showed a toxigenic strain is more sensitive to H_2O_2 than a non-toxigenic strain and suggest that this contradicts the model, but the model actually reproduces that experiment (our Fig. 3, right side). Huisman *et al*. also propose a different function for MC based on their more recent research, where MC enables acclimation to C-limitation. We don't question this mechanism and it would be useful to update the model to include it. Development of mechanistic models is a dynamic/stepwise process, and models will always lag biological understanding. There are many potential functions of MC and they are not mutually exclusive. What evidence is there to refute that MC (also) binds to proteins and protects them against oxidative damage, as it is implemented in the model?

Fifth, Huisman *et al*. suggest the ecology of toxigenic and non-toxigenic strains may be affected by competition for C, consistent with their more recently proposed biological role of MC (see above). The model, like most phytoplankton models, does not consider C-limitation, and we acknowledged above that it would be useful to extend the model. They also point to a study (*14*) that showed the toxigenic wildtype is a better competitor for N than the non-toxigenic mutant, and suggest this is opposite to the model. It is true that, in the model, the additional N required to make MC gives the toxigenic strain a disadvantage under N-limitation. However, although the mutant non-toxigenic strain used in that study is useful for exploring molecular mechanisms, it is not a good representative of wildtype non-toxigenic strains or their ecology. The deletion of the *mcyB* gene has consequences for the expression of many genes and it increases synthesis of other cyanopeptides and consequently N requirements (*14*). The model could reproduce those experiments, but it would require different parameterization of the two strains (beyond presence/absence of the *mcyB* gene) to reflect those differences.

Sixth, Huisman *et al*. criticize that the model only considers *Microcystis* and no other phytoplankton, and they suggest that nutrient reductions will lead to a shift away from *Microcystis* toward nontoxic eukaryotic phytoplankton. We acknowledge that any changes due to management or climate may result in a species shift, which would not be predicted by the model, and that adds uncertainty to our results. However, understanding and predicting species shifts is complicated, and it is not clear that nutrient reductions always or in the case of Lake Erie will lead to a shift toward eukaryotes. In Lake Erie (and many other systems), the present resurgence of *Microcystis* occurred following nutrient reductions (*15*). With the limited current understanding, assuming such a shift will not occur may be a good precautionary management approach.

Huisman *et al*. end their critique by pointing to the success of P load reductions in controlling cyanobacteria blooms, which is entirely consistent with the model (our Figs. 4B1 and 4C5) and misses the point of our paper, i.e., that it may increase toxin concentrations. The specific criticisms they provide are useful in that they point to uncertainties and potential future developments of the model. However, such uncertainties are unavoidable and inherent in all models of complex ecological systems. The model we presented is based on a large body of biological evidence, and even with its simplifications constitutes the most complete and consistent representation of *Microcystis* growth and toxin production available today. The model predicts that reducing P alone will increase toxin concentrations, which is in striking contrast to the present management approach, which assumes P load reductions will reduce toxicity. Management should be based on the best science and models available. We scientists need

[1]Water Quality Engineering, Technical University of Berlin, Berlin, Germany. [2]Department of Microbiology, University of Tennessee, Knoxville, TN. [3]Department of Earth and Environmental Science, University of Michigan, Ann Arbor, MA. [4]Cooperative Institute for Great Lakes Research, University of Michigan, Ann Arbor, MA.
***Corresponding author. Email: ferdi.hellweger@tu-berlin.de (F.L.H.); wilhelm@utk.edu (S.W.W.)**
†These authors contributed equally.

to be clear and maybe sometimes "provocative" to make this happen.

REFERENCES

1. F. L. Hellweger *et al*., *Science* **376**, 1001–1005 (2022).
2. G. Sandrini, S. Cunsolo, J. M. Schuurmans, H. C. P. Matthijs, J. Huisman, *Front. Microbiol.* **6**, 401 (2015).
3. D. B. Van de Waal *et al*., *Ecol. Lett.* **12**, 1326–1335 (2009).
4. K. Kameyama, N. Sugiura, Y. Inamori, T. Maekawa, *Environ. Toxicol.* **19**, 20–25 (2004).
5. C. O. P. Patterson, J. Myers, *Plant Physiol.* **51**, 104–109 (1973).
6. D. J. Smith *et al*., *Appl. Environ. Microbiol.* **88**, e0254421 (2022).
7. D. Smith, The Impact of Microbial Interactions and Hydrogen Peroxide on Western Lake Erie Cyanobacterial Blooms, University of Michigan (2021).
8. M. Drábková, H. C. P. Matthijs, W. Admiraal, B. Maršálek, *Photosynthetica* **45**, 363–369 (2007).
9. M. Drábková, W. Admiraal, B. Marsálek, *Environ. Sci. Technol.* **41**, 309–314 (2007).
10. E. F. J. Weenink *et al*., *Environ. Microbiol.* **23**, 2404–2419 (2011).
11. C. Dziallas, H.-P. Grossart, *PLOS ONE* **6**, e25569 (2011).
12. S. Meissner, J. Fastner, E. Dittmann, *Environ. Microbiol.* **15**, 1810–1820 (2013).
13. J. M. Schuurmans *et al*., *Harmful Algae* **78**, 47–55 (2018).
14. S. Suominen, V. S. Brauer, A. Rantala-Ylinen, K. Sivonen, T. Hiltunen, *Aquat. Ecol.* **51**, 117–130 (2017).
15. H. W. Paerl *et al*., *Hydrobiologia* **847**, 4359–4375 (2020).

ACKNOWLEDGMENTS

Funding: National Oceanographic and Atmospheric Administration grant NA18NOS4780175 (F.L.H. and S.W.W.). This is NOAA contribution 1034. National Institute of Environmental Health Sciences grant 1P01ES028939-01 (S.W.W.); National Science Foundation grant OCE-1840715 (S.W.W.); National Oceanographic and Atmospheric Administration grant NA17NOS4780186 (G.J.D.); National Science Foundation grant OCE-1736629 (G.J.D.); German Research Foundation [DFG, Research Training Group "Urban Water Interfaces (UWI)" (GRK 2032/1)]. **Author contributions:** Formal analysis: F.L.H., C.S., and F.E. Funding acquisition: F.L.H., S.W.W., and G.J.D. Project administration: F.L.H. and S.W.W. Software: F.L.H. Supervision: F.L.H., S.W.W. and G.J.D. Visualization: F.L.H. and C.S. Writing - original draft: F.L.H. Writing - review and editing: F.L.H., C.S., R.M.M., S.W.W., D.J.S., G.J.D., and F.E. **Competing interests:** Authors declare that they have no competing interests. **License information:** Copyright © 2022 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.sciencemag.org/about/science-licenses-journal-article-reuse

Submitted 18 July 2022; accepted 12 October 2022
10.1126/science.ade2277

By **Lan Nguyen Chaplin**

We are worthy

During the first meeting with my future mentor when I was an undergraduate student, he asked why I wanted to join his lab. I confidently responded with the first thing that came to mind: "Because you're famous." He let out the loudest laugh I have ever heard and told me students usually say they want to gain research experience to apply to grad school or they find his research fascinating. I wanted to hide under the table. I was so embarrassed I hadn't thought of a savvy answer like the other students. I didn't even know what grad school was, but it sounded fancy and therefore out of my league. These were some of my struggles as a first-generation college student and aspiring professional from an immigrant family.

I am the youngest of 14 children from a Vietnamese refugee family, and the American dream was very real to me, yet elusive at the same time. I knew college would be an important step and I was thrilled when I got into a good one close to home. But I knew virtually nothing about college and had no one to turn to for guidance. Everything was novel and intimidating, and I was always several steps behind, which chipped away at my confidence. I spent more time trying to figure out how to pay for college—navigating the work-study program, part-time jobs, grants, scholarships, financial aid forms—and wondering whether I should drop out than I did studying. My grades tanked. Someone like me wasn't meant to be where I was.



"I needed to believe I had value to offer … as a first-gen college student."

And so it was no wonder I blew that interview with Dr. D. I reached for my backpack to leave, thinking this meeting was over and so was my chance to work with this scientist. To my shock, after he finished laughing Dr. D welcomed me to his lab, telling me he liked my honesty. I worked in that lab for 2 years, and Dr. D helped lay a foundation for my confidence and self-worth. He taught me what it means to be a good mentor, especially to first-gen students.

From the start, Dr. D consistently asked me for help and encouraged me to share my ideas. At first I was stunned. What could I possibly have to offer? But despite my blank stares as he tried to extract ideas from me, he persisted, with uncanny patience. Sometimes he would even turn to work on something else to take the pressure off so I could work up the courage to offer a thoughtful response. He never doubted I had something worthwhile to contribute. Over time, I began to believe it, too.

He held me to the highest standards. Yet no matter what mistakes I made, he showed me grace and compassion—including sharing stories of his own missteps. He asked thoughtful questions about how my family celebrated Vietnamese Têt and how my bicultural upbringing shaped me, and he told me about his family in turn. He did not try to separate the person from the researcher, but instead sought to mentor my whole self.

When I was figuring out what to do after graduation, he told me to go where people needed me, appreciated my potential, and would happily put in the work to help me succeed. Beyond that, he added, I needed to believe I had value to offer and was worthy of support. That second part had been particularly difficult for me as a first-gen college student from an immigrant family, feeling I didn't belong and lacking the confidence to believe I could make a difference in the world. But Dr. D's persistent encouragement and support pushed me to accept it.

That was about 25 years ago. When I became a professor and began to mentor my own students, many of whom reminded me of my younger self, I developed a mentoring approach inspired by Dr. D. Its tenets are generosity, respect, authenticity, championing, and emboldening expectations—in a word, GRACE. With all my students, but especially those who are the first in their family to go to college, I strive to be generous with my time, compassion, expertise, and social capital; respect mentees' unique backgrounds and contributions; be an authentic human being first, then a professor; champion and advocate for my mentees whenever I can; and dare them to expect the best from themselves. As I say to my students, how can we expect others to see that we're worthy if we can't see it ourselves? ■

Lan Nguyen Chaplin is a professor at Northwestern University.
Send your career story to SciCareerEditor@aaas.org.

ILLUSTRATION: ROBERT NEUBECKER

PRIZE ESSAY

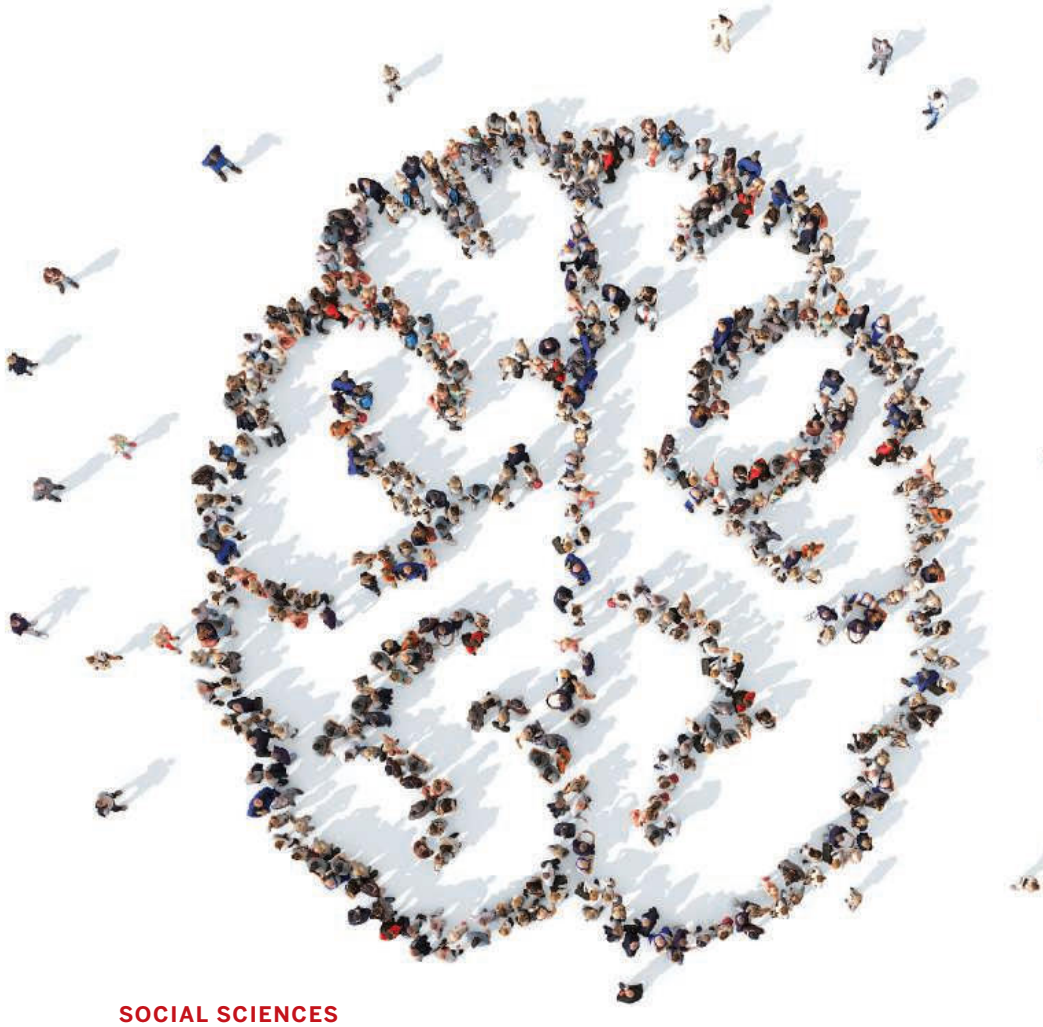
GRAND PRIZE WINNER

Bill Thompson



Bill Thompson received an undergraduate degree from Sheffield Hallam University and a PhD from the University of Edinburgh. After completing a postdoctoral fellowship at Princeton University, he started his lab in the Department of Psychology at the University of California, Berkeley, in 2022. His research examines the computational processes that support human reasoning, creativity, and language. **www.science.org/doi/10.1126/science.ade3128**

NOMIS & Science
YOUNG EXPLORER AWARD



SOCIAL SCIENCES

An ever-evolving mind

Experimental evolution of human cognition helps us understand ourselves

By **Bill Thompson**

In 1887, in an upstairs room in a secondary school in Sheffield, England, a fire broke out. The fire at Wesley College engulfed a custom-built incubator that had housed thousands of generations of a population of unicellular organisms. Every evening for 7 years, the incubator's temperature had been carefully increased by the diligent amateur scientist and school governor, William Henry Dallinger.

Dallinger had been studying how organisms adapt to changing conditions. By gradually manipulating the conditions inside the incubator over a long period of time, he had taken control of the selection pressures that influence how a species evolves. The accident destroyed one of the first known studies of experimental evolution—a high-cost, high-value experimental methodology that helps biologists test the predictions of evolutionary theory under controlled laboratory conditions.

I was born exactly 100 years later in 1987, less than a mile away in the University of Sheffield's Jessop Hospital for Women. I did not learn about Dallinger's experiments for another 31 years, but when I did, it changed my life and my career.

I am a cognitive scientist, and through my research, I try to understand what makes human intelligence so open-ended and creative. To accomplish this goal, I conduct my own high-cost, high-value experiments designed to grow new cognitive abilities in the laboratory, using experimental evolution.

Department of Psychology, University of California, Berkeley, Berkeley, CA, USA. Email: wdt@berkeley.edu

CREDITS: (PHOTO) COURTESY OF BILL THOMPSON; (ILLUSTRATI

In my experiments, nobody is in an incubator, and we do not turn up the heat. Instead, thousands of participants are asked to be as creative as they can when they face a novel psychological task, such as a problem-solving scenario, a decision-making dilemma, or an attempt to communicate a complex meaning without using words.

In my most ambitious studies (*1*), participants are organized into experimental generations. Each generation transmits their insights and discoveries to the next generation through use of language, demonstration, or teaching, for example. Over time, this creates an evolutionary process: The objects of evolution are not organisms—they are cognitive algorithms; and the mechanisms of transmission are not genes—they are thinking, talking people.

These experimental studies of cultural evolution help to test mathematical theories that I have developed to describe how human intelligence evolves through social interaction (*2*). Mathematical models and evolutionary experiments might sound like elaborate psychological methods, and this approach has certainly been a risky investment for me as a scientist. But I believe that meaningful progress in cognitive science depends on people taking this gamble, because until we push our science beyond its historical focus on individuals, our understanding of human cognition will always be incomplete.

The hypothesis guiding my research has deep roots in social psychology, anthropology, and linguistics. In my view, the social sciences have converged around a core principle of human intelligence, but the principle has been underappreciated in computational discourses surrounding human and artificial intelligence. The principle is that complex human cognitive functions such as language, self-awareness, and theory of mind arise not just from the neural architecture of individual brains but also from being embedded in larger processes of social interaction and cultural inheritance at the population level.

The goal of my research has been to translate this perspective into computational theories and experiments capable of testing them—a truly interdisciplinary challenge. When I tried to develop formal models of cognition as a PhD candidate in linguistics, I relied on methods from computer science, specifically, probabilistic machine learning, Bayesian inference, and stochastic optimization algorithms (*3*–*5*).

To extend these models to the population setting, I had to integrate them with mathematical models of evolutionary dynamics, in particular, replicator dynamics (*2*). These models helped to establish a formal integrative theory of intelligence in populations and revealed precise solutions to some long-standing mysteries surrounding nativism and empiricism.

As a postdoctoral researcher, I sought ways to translate the models I had developed into experiments that were able to test their predictions and to communicate their relevance to a broader audience. My postdoctoral mentor, Tom Griffiths, introduced me to the method of experimental evolution pioneered by Dallinger (*6*). Together, we worked to conceptualize an analogous framework applicable to cognitive science that scaled up experimental approaches to cultural evolution.

Experimental evolution in biology provided a model but not an implementation. Development of the necessary infrastructure to conduct experiments in multigenerational populations required years of deep investments of time and resources and collaboration with teams of software developers. In 2020, I used these new methods to conduct a massive experimental simulation of cultural evolution (*1*). This study of human problem-solving abilities showed how powerfully our cognitive algorithms can be shaped by social learning, which illuminated processes that occur in all societies and cultures. The experiment concluded a decade-long ambition to translate a long-standing idea from hypothesis to mathematical model, to computational theory, to a large-scale experimental test, and, ultimately, to publication of the results (*1*).

I trained in psychology as a first-generation undergraduate student in my hometown of Sheffield, where I lived in my parents' basement and worked multiple part-time jobs. In January 2022, I founded the Experimental Cognition Laboratory in the Department of Psychology at the University of California, Berkeley. With my students and collaborators, I will continue to pursue the basic science of cognition using experimental evolution. However, this journey has helped me to see that the technology that we developed creates a deeper opportunity. Experimental evolution in cognitive science can help us contribute to one of the newest and most-pressing challenges that faces modern information societies, because it provides a safe, ethical, transparent test ground for controlled scientific studies of social-networking algorithms and their impacts on thinking and reasoning in human populations.

In 1887, Dallinger showed that people can emulate nature by influencing selection on unicellular organisms. Today, we can substitute microorganisms with cognitive representations and substitute Dallinger's visits with complex social-networking algorithms. I hope that my work can help to pass on the cognitive algorithms that I have inherited from my mentors and Dallinger to a new generation of researchers who use creative and ambitious experimentation to build bridges between the biological, social, and cognitive sciences and to better understand the mechanisms shaping our minds and our societies. ■

REFERENCES AND NOTES

1. B. Thompson, B. van Opheusden, T. Sumers, T. L. Griffiths, *Science* **376**, 95 (2022).
2. B. Thompson, S. Kirby, K. Smith, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 4530 (2016).
3. B. Thompson, B. de Boer, *J. Lang. Evol.* **2**, 94 (2017).
4. M. Schouwstra, H. de Swart, B. Thompson, *Cogn. Sci.* **43**, e12732 (2019).
5. B. Thompson, T. L. Griffiths, *Proc. R. Soc. London Ser. B* **288**, 20202752 (2021).
6. W. H. Dallinger, *Proc. R. Soc. London* **27**, 332 (1878).

10.1126/science.ade3128

FINALIST

Célia Lacaux



Célia Lacaux received her undergraduate degree from Aix-Marseille University and Imperial College London and her master's degrees from Ecole Normale Supérieure and University College London. She then completed a PhD in 2021 at the Paris Brain Institute under the supervision of D. Oudiette and I. Arnulf. She is now moving to the lab of S. Schwartz in Switzerland for a postdoctoral position. Her research focuses primarily on the relationship between sleep and creativity. **www.science.org/doi/10.1126/science.ade3129**

FINALIST

Stephen Kissler



Stephen Kissler received his undergraduate degree from the University of Colorado and a PhD from the University of Cambridge, both in applied mathematics. He is completing his postdoctoral fellowship at the Harvard T.H. Chan School of Public Health and will start his lab in the Department of Computer Science at the University of Colorado in 2023. His research examines how immunological and behavioral factors influence the spread of respiratory viruses. **www.science.org/doi/10.1126/science.ade3133**

PHOTOS: (TOP TO BOTTOM) COURTESY OF CÉLIA LACAUX; KENT DAYTON

PRIZE ESSAY

FINALIST

Célia Lacaux



Célia Lacaux received her undergraduate degree from Aix-Marseille University and Imperial College London and her master's degrees from Ecole Normale Supérieure and University College London. She then completed a PhD in 2021 at the Paris Brain Institute under the supervision of Delpine Oudiette and Isabelle Arnulf. She is now moving to the lab of Sophie Schwartz in Switzerland for a postdoctoral position. Her research focuses primarily on the relationship between sleep and creativity.

www.science.org/doi/10.1126/science.ade3129

NOMIS & Science
YOUNG EXPLORER AWARD

SOCIAL SCIENCES

A doorway into possibility

The borderland between wakefulness and sleep promotes creativity

By **Célia Lacaux**

Each night, we cross a bridge that connects the waking and sleeping worlds. We know very little about this bridge that symbolizes the sleep-onset period because our passage is brief and leaves only a few fragmented memories behind. Moreover, sleep researchers have largely overlooked this twilight period, likely because of its "in-between" and fleeting nature. However, upon closer examination, the sleep-onset period appears to be a rich and dynamic time during which our body and mind undergo substantial changes. Brain activity slows, muscles relax, and reality gradually distorts as dreamlike images begin to dance before the eyelids.

In contrast to sleep researchers, many artists and inventors, such as Thomas Edison and Salvador Dali, have been fascinated by this period and viewed it as a real source of creative inspiration. They even devised methods for capturing this fleeting moment and its flashes of creativity before they vanished into the limbo of sleep. Their secret? Taking naps while holding an object that dropped noisily as they dozed off to awaken them just in time to record some of their discoveries and/or ideas. Is there any truth in this alluring story? Is the sleep-onset period conducive to creativity?

This question, at the intersection of the life and social sciences, was the focus of my PhD thesis work. My main hypothesis was that hybrid states, at the borderland between wakefulness and sleep, promote creativity. To test this hypothesis, I examined a physiological state in which sleep and wakefulness coexist (the sleep-onset period) alongside a sleep disorder, narcolepsy, in which the line between these two states is finer than usual.

Patients with narcolepsy experience excessive daytime sleepiness, which leads to recurrent bouts of sleep during the day. These individuals are also dream "maestros": They remember their dreams more frequently than do the general population and are predominantly lucid dreamers (they are conscious of dreaming while in a dream and can even control their dreams' scenarios)—two factors that have been positively correlated with creativity (*1*, *2*). Could narcoleptic individuals' atypical experience with sleep onset and dreams promote the development of increased creativity over time? To address this question, I administered various questionnaires and creativity tests (for example, invent different endings to a story) to 185 patients with narcolepsy and 126 healthy subjects (*3*). Patients with narcolepsy scored higher than did healthy subjects on all tests. This higher creative potential also correlated positively with indices of sleepiness and with symptoms of narcolepsy indicative of a hybrid state between wakefulness and sleep (such as hypnagogic hallucinations, sleep paralysis, or lucid dreaming). These findings demonstrate the existence of an increased creative potential in narcolepsy. In addition to providing a silver lining to a disabling disorder, these results pointed to a possible link between hybrid states and creativity.

I decided to go further and directly test the impact on creativity of a hybrid physiological state and focused on the sleep-onset period, commonly known as N1. To do so, I took an approach to empirically test the method used by Thomas Edison to increase his creativity. In this second study, 103 participants were asked to solve lengthy and tedious mathematical problems without knowing that a hidden rule could allow them to solve the problems almost instantly (*4*). Participants had to compute many of these arithmetic problems before and after a 20-minute break. The break was carried out under the same conditions as those used by Thomas Edison: The subjects sat in a chair and were instructed to rest while holding an object in one hand. If the object fell, they had to report everything that came into their heads before the object fell. During the break, participants' brain activity was monitored to determine their wake-sleep stages.

My collaborators and I found that after an average of only one minute in N1, most subjects (83%) suddenly discovered the hidden rule—a threefold increase in the proportion of individuals experiencing this insight as compared with subjects who had remained awake during the break. This beneficial effect disappeared as the sub-

Institut du Cerveau–Paris Brain Institute, Pitié Salpêtrière University Hospital, Paris, France.
Email: celia.lacaux@gmail.com

PHOTO: COURTESY OF CÉLIA LACAUX

jects reached deeper sleep. We also found evidence of a creative brain signature by analyzing further the neuronal factors that could predict these creative bursts. The probability of problem-solving was greatest when the subjects' brain waves contained a moderate amount of alpha (a marker of the wake-sleep transition) and a low level of delta (a marker of deep sleep) oscillations. We revealed the existence of a creative optimum within the sleep-onset period, and to reach this spot, one must balance falling asleep easily against falling asleep too deeply. Edison's technique proved effective in keeping subjects in this "in-between" state by preventing them from falling completely asleep. It was also successful at capturing dreams that occur during sleep onset—a technique that could be used in the future to explore the dreaming world and search for its brain signature.

This work lies at the crossroads of clinical and fundamental research, bridging diverse disciplines (sociology, psychology, medicine, and neuroscience) and research fields (such as sleep and creativity). My results show that sleep onset represents a doorway into creativity, a finding with substantial societal implications. Creativity is indeed one of the most essential human abilities. Without it, we would live in a world devoid of, for example, art, the internet, vaccines, or spaceships. Yet despite its utter importance in our lives, we appear to have no control over creativity—ideas seem to come up ex nihilo. By revealing the existence of a creative sweet spot, my work offers everyone an opportunity to summon their creative muse at will if they use Edison's technique during their own micro-naps. ■

REFERENCES AND NOTES

1. R. Vallat, B. Türker, A. Nicolas, P. Ruby, *Nat. Sci. Sleep* **14**, 265 (2022).
2. N. Zink, R. Pietrowsky, *Int. J. Dream Res.* **6**, 98 (2013).
3. C. Lacaux *et al.*, *Brain* **142**, 1988 (2019).
4. C. Lacaux *et al.*, *Sci. Adv.* **7**, eabj5866 (2021).

10.1126/science.ade3129

PRIZE ESSAY

FINALIST

Stephen Kissler



Stephen Kissler received his undergraduate degree from the University of Colorado and a PhD from the University of Cambridge, both in Applied Mathematics. He is completing his postdoctoral fellowship at the Harvard T.H. Chan School of Public Health and will start his lab in the Department of Computer Science at the University of Colorado in 2023. His research examines how immunological and behavioral factors influence the spread of respiratory viruses. **www.science.org/doi/10.1126/science.ade3133**

NOMIS & Science
YOUNG EXPLORER AWARD

SOCIAL SCIENCES

Revealing contagion

Mathematical models help predict and manage the course of pandemics

By **Stephen Kissler**

I feel a kinship with the artists and poets of the Middle Ages who tried to make sense of the plague. They left us chilling, yet oddly familiar images of dancing skeletons playing flutes, that summoned contagion up from the graves; images of stars detached from the sky that rained down illness—which captured beliefs that were half science and half dream (*1, 2*). These individuals left us words that our modern lexicons have failed to replace: influenza, from the influence of the stars; cholera, from an excess of the fiery humor; syphilis, from the name of an epic poem's unfortunate hero (*3*). Through their images and words, these artists and authors turned an invisible microbe into something that could be seen, heard, and somehow managed.

Mathematics, like art and poetry, lets us see the unseeable. We can use mathematics to trace the courses of planets, predict devastating storms, and learn the properties of subatomic particles through the traces they leave behind. Likewise, and in the tradition of the plague artists, we can trace the global course of an illness, predict when it will surge and recede, and learn the properties of a virus through the imprint it leaves on the body. Typically, the mathematics of contagion rest on a sure scientific foundation, but sometimes we are thrown into the turmoil of grappling with something completely new. It was into this turmoil that my colleagues and I were thrown when we heard the first reports of an atypical pneumonia in Wuhan, China, on 4 January 2020.

The early reports from Wuhan and Lombardy, Italy, made clear that we needed a swift response—but how? What would it take to bend the epidemic's trajectory? Without a clear frame of reference, we relied on metaphor: Was this the outbreak of 1918, 2003, or 2009 (*4–6*)? The right comparison was key, yet despite four other coronaviruses spreading regularly beneath our literal noses, we knew little about them. We dug through historical records and gathered what data we could. Then, we built upon our field's fundamental equations, to construct a mathematical model to explain how these common coronaviruses behave (*7*). We found that they are powerfully driven by seasonal shifts in our behavior, they can induce immunity against one another, and yet our immunity to them rapidly wanes. Based upon these insights, we built a series of best guesses for how the novel coronavirus' trajectory might unfold: a major surge with recurring outbreaks, and years of intermittent control measures needed to keep hospitals from being overwhelmed (*8*). The equations made clear that the pandemic of a lifetime was at our doorstep, and it would take all of us to shift its course.

Despite these predictions, we were unprepared for how capriciously the virus would strike, like a cyclone flattening neighborhoods at random. The only clear pattern that emerged was that the new virus showed a predilection for the poor and the marginalized (*9*). Was this simply due to differences in underlying health, or did some communities also experience much higher rates of infection? At the time, we had little idea where the virus was as barriers to testing had obscured its movements (*10*). We needed the epidemiological equivalent of weather stations to map and predict its path (*11*). Fortunately, a network of hospitals across New York City provided just this, by routinely administering severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) tests to women who were about to give birth (*12*). We used a new statistical approach to triangulate the level of disease circulating in these women's communities and found vast differences, which reflected the disparities in severe disease occurrence (*13, 14*). We then sought to explain what drove these disparities. A database of anonymized cell phone locations over time was used, and we found that the neighborhoods with the most infections also had the most commuters (i.e., frontline workers—those who were regularly still commuting from home to work—were bearing the brunt of the pandemic). We used these results to advocate for greater protections for essential workers.

Amid the pandemic's march across the globe and into our communities, a parallel drama played out between the virus and our bodies. An infection is its own small

Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA, USA. Email: skissler@hsph.harvard.edu

PHOTO: KENT DAYTON

epidemic, where the virus surges, peaks, and recedes as it is beaten back by our immune cells. As this process repeated, the virus began to change. Small pieces of the virus's script for replication were deleted and switched until suddenly it became more contagious and more deadly. But we changed, too; vaccines trained our bodies to recognize the virus and subdue it more quickly. Our communities had to adapt, with new policies for quarantine and isolation, better guidance for masking, and clearer communication of the risks. To provide this guidance, we needed a clear picture of the virus's course during an infection, and how this differed by variant and vaccination status.

An opportunity arose when a major sports league began testing their players and staff daily to reduce the risk of outbreaks. We utilized a model for the virus's struggle with our immune system and used their data to distinguish the effect of the variant from the effect of the vaccine. Vaccinated people cleared their infections more quickly, which suggested they did not need to isolate for as long as those who were unvaccinated. But surprisingly, we found few differences between variants: All produced similar amounts of virus within the body for similar amounts of time (*15*). This finding left only one plausible explanation for the increased contagiousness of the new variants: A stronger bond to the surfaces of our cells. As we charted the virus's course through the body, we could peer indirectly at the structure of the virus itself.

Despite our recent major advances in science and medicine, this pandemic has caused me to wonder whether we are perhaps not so different from the plague artists of centuries ago. We experience similar fears, similar wonder, and an equally relentless desire to know. We create images and write stories that reflect a marriage between our imagination and reality, and through this we provide some insight, some guidance, and some hope. The medium has changed, from woodcuts to equations, but the aim is no different: We seek only to glimpse that which cannot be seen. ■

REFERENCES AND NOTES

1. H. Schedel, *Liber Chronicarum*, Wellcome Collection, 1493; https://wellcomecollection.org/works/h5z6hxqg/images?id=ex69jewk.
2. K. Lykosthenes, *Prodigiorum Ac Ostentorum Chronicon, Quae Praeter Naturae Ordinem, Motum, et Operationem, et in Superioribus & His Inferioribus Mundi Regionibus*, Wellcome Collection, 1557; https://wellcomecollection.org/works/kqd9eqrg/images?id=nyxuv4nk.
3. G. Fracastoro, *Syphilis, Sive Morbus Gallicus*, 1530.
4. D. M. Morens, A. S. Fauci, *J. Infect. Dis.* **195**, 1018 (2007).
5. M. Lipsitch *et al.*, *Science* **300**, 1966 (2003).
6. J. M. Wood, *Influenza Other Respir. Viruses* **3**, 197 (2009).
7. W. O. Kermack, A. G. McKendrick, *Proc. R. Soc. A Math. Phys. Eng. Sci.* **115**, 700 (1927).
8. S. M. Kissler, C. Tedijanto, E. Goldstein, Y. H. Grad, M. Lipsitch, *Science* **368**, 860 (2020).
9. J. T. Chen, N. Krieger, *J. Public Health Manag.* **27**, S43 (2021).
10. S. Kissler, "Let's finally get COVID-19 testing right," *The Hill*, 25 May 2021.
11. C. Rivers, D. George, "How to Forecast Outbreaks and Pandemics," *Foreign Affairs*, 29 June 2020.
12. D. Sutton, K. Fuchs, M. D'Alton, D. Goffman, *N. Engl. J. Med.* **382**, 2163 (2020).
13. D. B. Larremore *et al.*, *eLife* **10**, e64206 (2021).
14. S. M. Kissler *et al.*, *Nat. Commun.* **11**, 4674 (2020).
15. S. M. Kissler *et al.*, *N. Engl. J. Med.* **385**, 2489 (2021).

10.1126/science.ade3133