How dimming city lights can save migrating birds p. 340

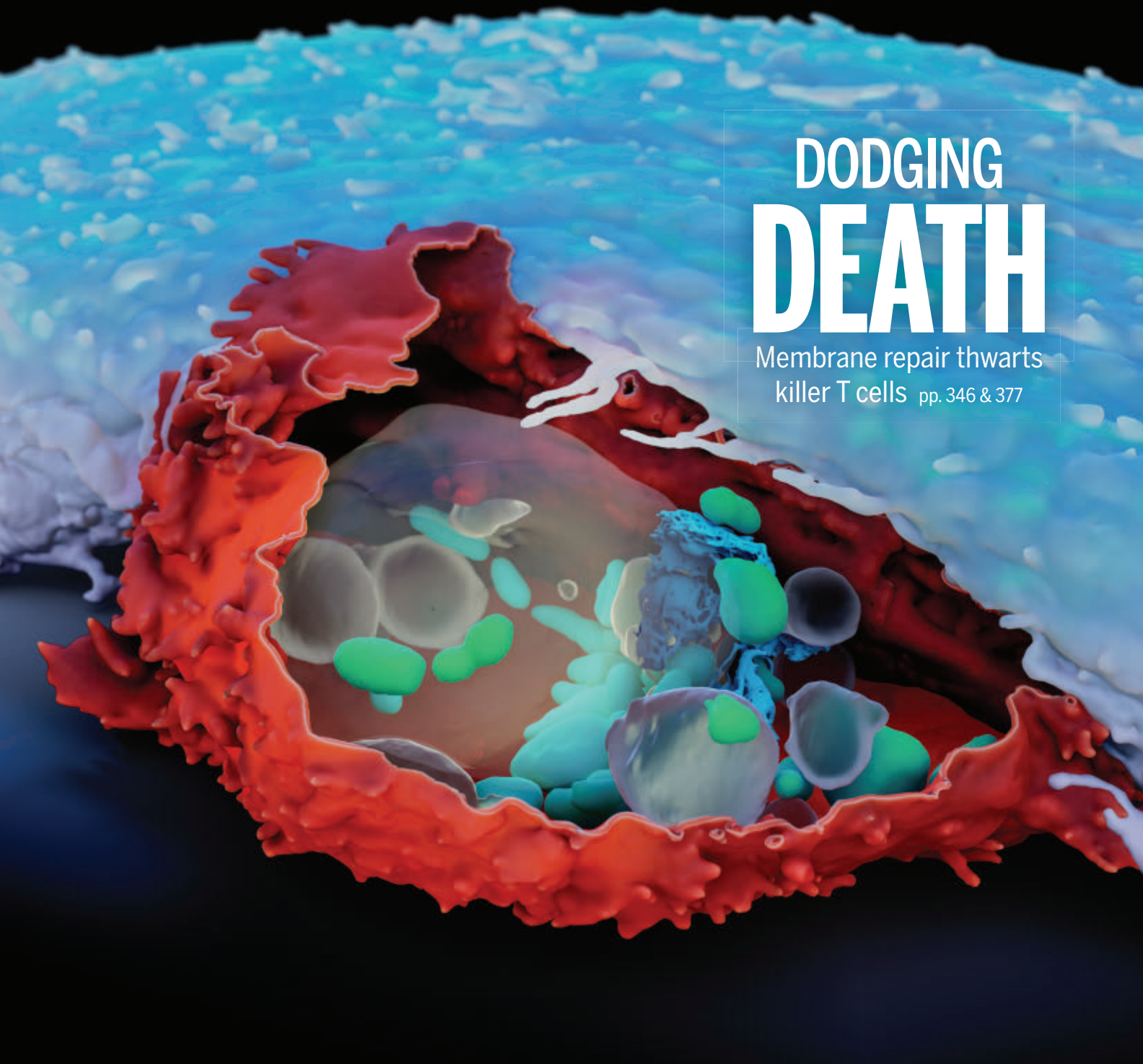
Evolved responses to fishing in Atlantic salmon pp. 344 & 420

Single-step hydrogen purification pp. 348 & 390

Science

$15
22 APRIL 2022
science.org

AAAS



DODGING DEATH

Membrane repair thwarts killer T cells pp. 346 & 377

WE'RE CELEBRATING.

A NEW ANTIBODY FAILED TO FAIL.

The US FDA has granted a Breakthrough Therapy designation to the first anti-TIGIT therapy for use in NSCLC in combination with PD-L1 blockade. TIGIT expression in paraffin-embedded non-small cell lung cancer is observed using TIGIT (E5Y1W) XP® Rabbit mAb #99567.

OUR PROCESS — YOUR SUCCESS.

As a privately held company founded and led by active research scientists, we understand your needs as a researcher. Like you, we only want to use products that are specific and deliver reproducible results. We validate every one of our thousands of antibodies and research products in-house. If any product fails to meet our stringent standards, we won't sell it. But when one passes, it's a big deal. For over 20 years, researchers have trusted CST to help them achieve experimental success. You can, too.

www.cellsignal.com

Cell Signaling TECHNOLOGY®

© 2022 Cell Signaling Technology, Inc. Cell Signaling Technology, CST, and XP are trademarks or registered trademarks of Cell Signaling Technology, Inc. 22-BCH-13880-MAR22

The World Laureates Association Prize (the " WLA Prize ") is established to support gobal science and technology advancement, address the challenges to mankind and promote the long-term progress of society.

The 2022 WLA Prize Call for Nominations

Championing excellence in science

The inaugural World Laureates Association Prize (the " WLA Prize ") is now open for nominations.

The WLA is a non-profit, non-governmental international organization. The membership includes laureates of the most prestigious science awards, including Nobel Prize, Turing Award, and Fields Medal.

Nominations are being accepted for the **" WLA Prize in Computer Science or Mathematics "** and the **" WLA Prize in Life Science or Medicine "**. Each Prize is for a total of **RMB 10 Million**. The deadline for submission of nominations is **June 1st, 2022**.

All nomination information can be found at **www.wlaprize.org**.

NOMIS & Science
YOUNG EXPLORER AWARD 2022

Research at the intersection of the social and life sciences

Unconventional. Interdisciplinary. Bold.

The **NOMIS & *Science* Young Explorer Award** recognizes and rewards early-career M.D., Ph.D., or M.D./Ph.D. scientists that perform research at the intersection of the social and life sciences. Essays written by these bold researchers on their recent work are judged for clarity, scientific quality, creativity, and demonstration of cross-disciplinary approaches to address fundamental questions.

A cash prize of up to 15 000 USD will be awarded to essay winners, and their engaging essays will be published in ***Science***. Winners will also be invited to share their work and forward-looking perspective with leading scientists in their respective fields at an award ceremony as well as a meeting of the NOMIS Board of Directors to consider future funding.

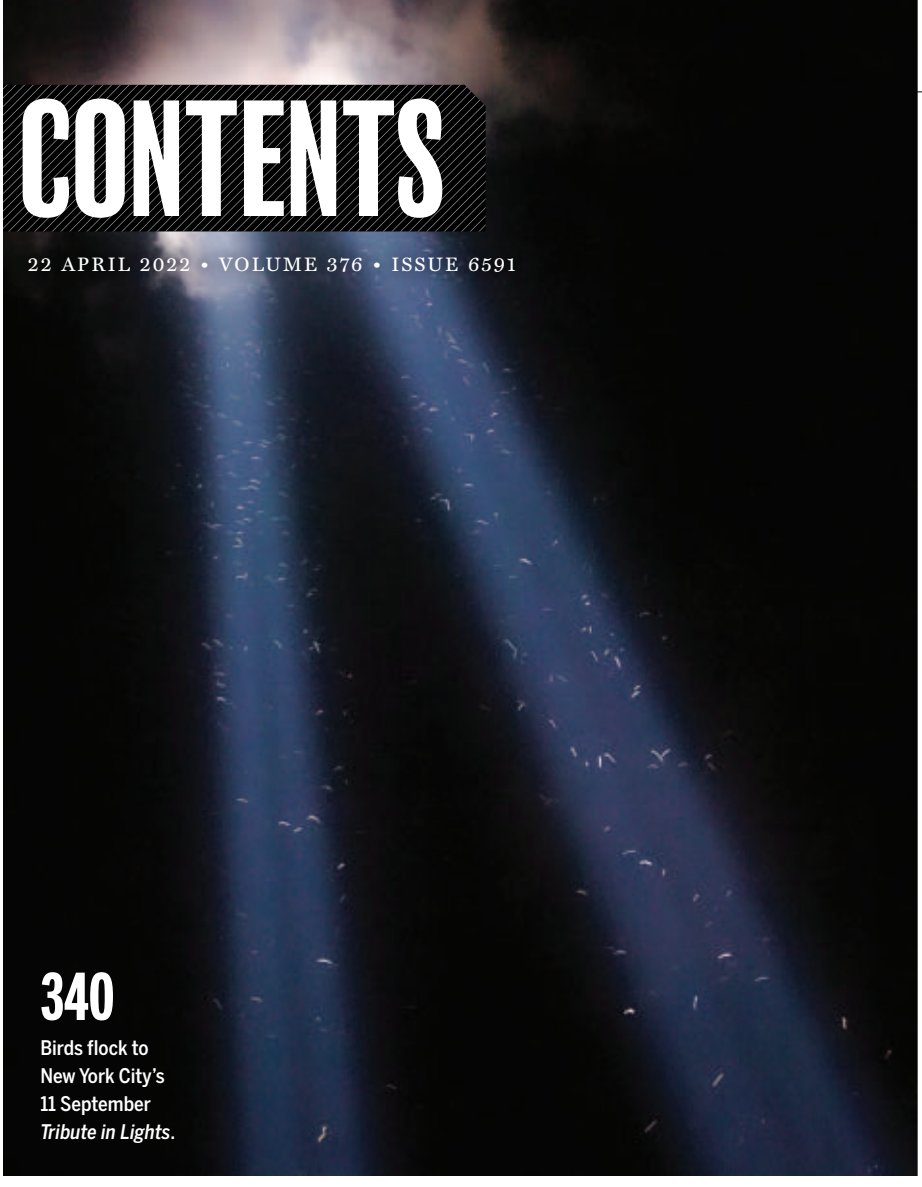
Apply by May 15, 2022
at www.science.org/nomis

NOMIS FOUNDATION
Creating the Spark

Science
AAAS

CONTENTS

22 APRIL 2022 • VOLUME 376 • ISSUE 6591



340
Birds flock to New York City's 11 September *Tribute in Lights*.

PHOTO: LAWRENCE LUCIER/FILMMAGIC

NEWS

IN BRIEF

330 News at a glance

IN DEPTH

332 Uranus should be NASA's top target, report finds
Influential survey recommends $4.2 billion flagship mission to the ice giant *By P. Voosen*

333 China refuses to end harsh lockdowns
Still not fully prepared for living with the virus, the government is sticking to "zero COVID" *By D. Normile*

335 Pterosaurs were clad in colorful plumage
Study suggests feathers arose—and were used for display—well before reign of dinosaurs *By R. Pérez Ortega*

336 Fighting cancer mutations by ... making more mutations?
Chemotherapies that induce new DNA errors may help unleash the immune system on tumors *By J. Kaiser*

337 Aid keeps researchers afloat in war-torn Ukraine
Fledgling programs aim to sustain lives and work, and stem brain drain *By R. Stone*

338 Dutch dangle bid for gravitational wave detector
Government pledges to support nearly half the cost of the Einstein Telescope *By E. Cartlidge*

FEATURES

340 Bright lights, big pity
City lights lure migratory birds, with lethal results, but weather radar and modeling can help reduce toll *By J. Sokol*
PODCAST

INSIGHTS

PERSPECTIVES

344 A long evolutionary reach for fishing nets
Earlier maturation of Atlantic salmon is linked to indirect effects of fisheries on its prey *By N. O. Therkildsen and M. L. Pinsky*
REPORT p. 420

346 Resisting attack by repairing the damage
Removing membrane pores may help cancer cells survive T cell assault *By N. W. Andrews*
RESEARCH ARTICLES p. 377

347 How safe are safety messages?
Highway fatalities increased in response to certain messages *By G. Ullman and S. Chrysler*
RESEARCH ARTICLES p. 370

348 Electrifying membranes to deliver hydrogen
An electrochemical membrane reactor enables efficient hydrogen generation *By A. J. Shih and S. M. Haile*
REPORT p. 390

350 Superconductors gain momentum
Spin-density modulations point to inhomogeneous superconductivity in a perovskite *By E. Pavarini*
REPORT p. 397

351 A fresh look at somatic mutations in cancer
Analysis of cancer genome sequences reveals new mutational signatures *By D. Szüts*
RESEARCH ARTICLES p. 368

353 Repetitive DNA in disease
Transposons become a focus of speculation and scrutiny in biomedical research *By K. H. Burns*

POLICY FORUM

355 Countering the future chemical weapons threat
Shift focus from disarmament to preventing reemergence *By T. H. Nguyen*

BOOKS *ET AL.*

359 Animal agriculturalists
There is much to learn from the farming practices of ants, termites, and other insects *By J. C. Postigo and J. E. Strassmann*

360 Inside the dinosaurs' demise
A science writer offers an accessible account of the Mesozoic's mass extinction *By V. Arbour*

Science
JOURNALS AAAS

Publish your research in the *Science* family of journals

The *Science* family of journals (*Science, Science Advances*, *Science Immunology, Science Robotics, Science Signaling*, and *Science Translational Medicine)* are among the most highly-regarded journals in the world for quality and selectivity. Our peer-reviewed journals are committed to publishing cutting-edge research, incisive scientific commentary, and insights on what's important to the scientific world at the highest standards.

Submit your research today!

Learn more at **Science.org/journals**

LETTERS

361 Offshore renewables need an experimental mindset
By A. F. Johnson et al.

361 Green energy threatens Chile's Magallanes Region
By H. V. Norambuena et al.

362 Brazilian pesticides law could poison the world
By L. Carneiro et al.

RESEARCH

IN BRIEF

364 From *Science* and other journals

REVIEW

367 Metamaterials
Tunable structured light with flat optics
A. H. Dorrah and F. Capasso
REVIEW SUMMARY; FOR FULL TEXT: DOI.ORG/10.1126/SCIENCE.ABI6860

RESEARCH ARTICLES

368 Cancer genomics
Substitution mutational signatures in whole-genome–sequenced cancers in the UK population
A. Degasperi et al.
RESEARCH ARTICLE SUMMARY; FOR FULL TEXT: DOI.ORG/10.1126/SCIENCE.ABL9283
PERSPECTIVE p. 351

369 Coronavirus
Structural basis for potent antibody neutralization of SARS-CoV-2 variants including B.1.1.529
T. Zhou et al.
RESEARCH ARTICLE SUMMARY; FOR FULL TEXT: DOI.ORG/10.1126/SCIENCE.ABN8897

370 Traffic safety
Can behavioral interventions be too salient? Evidence from traffic safety messages
J. D. Hall and J. M. Madsen
RESEARCH ARTICLE SUMMARY; FOR FULL TEXT: DOI.ORG/10.1126/SCIENCE.ABM3427
PERSPECTIVE p. 347

371 Electrooptics
Ferroelectric crystals with giant electro-optic property enabling ultracompact Q-switches
X. Liu et al.

377 Cell biology
ESCRT-mediated membrane repair protects tumor-derived cells against T cell attack
A. T. Ritter et al.
PERSPECTIVE p. 346

383 Protein design
Computational design of mechanically coupled axle-rotor protein assemblies
A. Courbet et al.

REPORTS

390 Catalysis
Single-step hydrogen production from NH_3, CH_4, and biogas in stacked proton ceramic reactors *D. Clark* et al.
PERSPECTIVE p. 348

394 Cell biology
Cell division in tissues enables macrophage infiltration *M. Akhmanova* et al.

397 Superconductivity
Superconducting spin smecticity evidencing the Fulde-Ferrell-Larkin-Ovchinnikov state in Sr_2RuO_4 *K. Kinjo* et al.
PERSPECTIVE p. 350

401 Quantum sensing
Atomic-scale quantum sensing based on the ultrafast coherence of an H_2 molecule in an STM cavity *L. Wang* et al.

406 Optoelectronics
Optical absorption of interlayer excitons in transition-metal dichalcogenide heterostructures *E. Barré* et al.

410 Organic chemistry
Controlling Ni redox states by dynamic ligand exchange for electroreductive Csp3–Csp2 coupling *T. B. Hamby* et al.

416 Solar cells
Organometallic-functionalized interfaces for highly efficient inverted perovskite solar cells *Z. Li* et al.

420 Fisheries
Rapid evolution in salmon life history induced by direct and indirect effects of fishing
Y. Czorlich et al.
PERSPECTIVE p. 344

DEPARTMENTS

329 Editorial
Africa's growing space enterprise
By T. Aganaba and E. Offiong

426 Working Life
Nurturing my ambitions *By K. F. Boehnke*

ON THE COVER

A killer T cell (red) attacks an ovarian cancer cell (white) in this digital reconstruction of high-resolution electron micrographs. A view into the T cell reveals the positions of its internal structures (organelles) at the precise moment that the T cell attempts to destroy its target. The cancer cell, however, resists this assault by activating a mechanism to seal perforations in its plasma membrane. See pages 346 and 377. *Data: Alex Ritter, Genentech, Inc.; Gleb Shtengel and C. Shan Xu, Janelia Research Campus; Rendering: Sebastian Spaar, ariadne.ai*

Science Staff 328
Science Careers 424

SCIENCE (ISSN 0036-8075) is published weekly on Friday, except last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005. Periodicals mail postage (publication No. 484460) paid at Washington, DC, and additional mailing offices. Copyright © 2022 by the American Association for the Advancement of Science. The title SCIENCE is a registered trademark of the AAAS. Domestic individual membership, including subscription (12 months): $165 ($74 allocated to subscription). Domestic institutional subscription (51 issues): $2212; Foreign postage extra: Air assist delivery: $98. First class, airmail, student, and emeritus rates on request. Canadian rates with GST available upon request, GST #125488122. Publications Mail Agreement Number 1069624. **Printed in the U.S.A.**
Change of address: Allow 4 weeks, giving old and new addresses and 8-digit account number. **Postmaster:** Send change of address to AAAS, P.O. Box 96178, Washington, DC 20090–6178. **Single-copy sales:** $15 each plus shipping and handling available from backissues.science.org; bulk rate on request. **Authorization to reproduce** material for internal or personal use under circumstances not falling within the fair use provisions of the Copyright Act can be obtained through the Copyright Clearance Center (CCC), www.copyright.com. The identification code for *Science* is 0036-8075. *Science* is indexed in the *Reader's Guide to Periodical Literature* and in several specialized indexes.

PHOTO: MICHEL ROGGO/MINDEN PICTURES

Science AAAS

Editor-in-Chief Holden Thorp, hthorp@aaas.org

Executive Editor Monica M. Bradford

Editors, Research Valda Vinson, Jake S. Yeston **Editor, Insights** Lisa D. Chong

DEPUTY EDITORS Stella M. Hurtley (UK), Phillip D. Szuromi, Sacha Vignieri **SR. EDITORIAL FELLOW** Andrew M. Sugden (UK) **SR. EDITORS** Gemma Alderton (UK), Caroline Ash (UK), Brent Grocholski, Pamela J. Hines, Di Jiang, Priscilla N. Kelly, Marc S. Lavine (Canada), Yevgeniya Nusinovich, Ian S. Osborne (UK), Beverly A. Purnell, L. Bryan Ray, H. Jesse Smith, Keith T. Smith (UK), Jelena Stajic, Peter Stern (UK), Valerie B. Thompson, Brad Wible, Yuen Yiu **ASSOCIATE EDITORS** Michael A. Funk, Bianca Lopez, Seth Thomas Scanlon (UK), Yury V. Suleymanov, Ekeoma Uzogara **LETTERS EDITOR** Jennifer Sills **LEAD CONTENT PRODUCTION EDITORS** Harry Jach, Lauren Kmec **CONTENT PRODUCTION EDITORS** Amelia Beyna, Jeffrey E. Cook, Chris Filiatreau, Julia Haber-Katris, Nida Masiulis, Abigail Shashikanth, Suzanne M. White **SR. EDITORIAL COORDINATORS** Carolyn Kyle, Beverly Shields **EDITORIAL COORDINATORS** Aneera Dobbins, Joi S. Granger, Jeffrey Hearn, Lisa Johnson, Maryrose Madrid, Ope Martins, Shannon McMahon, Jerry Richardson, Hilary Stewart (UK), Alice Whaley (UK), Anita Wynn **PUBLICATIONS ASSISTANTS** Alexander Kief, Ronmel Navas, Isabel Schnaidt, Brian White **EXECUTIVE ASSISTANT** Jessica Slater **ASI DIRECTOR, OPERATIONS** Janet Clements (UK) **ASI SR. OFFICE ADMINISTRATOR** Jessica Waldock (UK)

News Editor Tim Appenzeller

NEWS MANAGING EDITOR John Travis **INTERNATIONAL EDITOR** Martin Enserink **DEPUTY NEWS EDITORS** Elizabeth Culotta, Lila Guterman, David Grimm, Eric Hand (Europe), David Malakoff **SR. CORRESPONDENTS** Daniel Clery (UK), Jon Cohen, Jeffrey Mervis, Elizabeth Pennisi **ASSOCIATE EDITORS** Jeffrey Brainard, Kelly Servick, Catherine Matacic **NEWS REPORTERS** Adrian Cho, Jennifer Couzin-Frankel, Jocelyn Kaiser, Rodrigo Pérez Ortega (Mexico City), Robert F. Service, Erik Stokstad, Paul Voosen, Meredith Wadman **INTERN** Tess Joose **CONTRIBUTING CORRESPONDENTS** Warren Cornwall, Andrew Curry (Berlin), Ann Gibbons, Sam Kean, Eli Kintisch, Kai Kupferschmidt (Berlin), Andrew Lawler, Mitch Leslie, Eliot Marshall, Virginia Morell, Dennis Normile (Tokyo), Elisabeth Pain (Careers), Charles Piller, Gabriel Popkin, Michael Price, Joshua Sokol, Richard Stone, Emily Underwood, Gretchen Vogel (Berlin), Lizzie Wade (Mexico City) **CAREERS** Rachel Bernstein (Editor), Katie Langin (Associate Editor) **COPY EDITORS** Julia Cole (Senior Copy Editor), Morgan Everett, Cyra Master (Copy Chief) **ADMINISTRATIVE SUPPORT** Meagan Weiland

Creative Director Beth Rakouskas

DESIGN MANAGING EDITOR Marcy Atarod **GRAPHICS MANAGING EDITOR** Chris Bickel **INTERIM MULTIMEDIA MANAGING PRODUCER** Sarah Crespi **PHOTOGRAPHY MANAGING EDITOR** William Douthitt **WEB STRATEGY MANAGER** Kara Estelle-Powers **DESIGN EDITOR** Chrystal Smith **DESIGNER** Christina Aycock **GRAPHICS EDITOR** Nirja Desai **INTERACTIVE GRAPHICS EDITOR** Kelly Franklin **SENIOR GRAPHICS SPECIALISTS** Holly Bishop, Nathalie Cary **SENIOR SCIENTIFIC ILLUSTRATOR** Valerie Altounian **SCIENTIFIC ILLUSTRATOR** Ashley Mastin **VIDEO PRODUCER** Meagan Cantwell **SENIOR PHOTO EDITOR** Emily Petersen **PHOTO EDITOR** Kaitlyn Dolan **SOCIAL MEDIA STRATEGIST** Jessica Hubbard **SOCIAL MEDIA PRODUCER** Sabrina Jenkins **WEB DESIGNER** Jennie Pajerowski

Chief Executive Officer and Executive Publisher Sudip Parikh

Publisher, Science Family of Journals Bill Moran

DIRECTOR, BUSINESS SYSTEMS AND FINANCIAL ANALYSIS Randy Yi **DIRECTOR, BUSINESS OPERATIONS & ANALYSIS** Eric Knott **DIRECTOR OF ANALYTICS** Enrique Gonzales **MANAGER, BUSINESS OPERATIONS** Jessica Tierney **MANAGER, BUSINESS ANALYSIS** Cory Lipman **BUSINESS ANALYSTS** Kurt Ennis, Maggie Clark **FINANCIAL ANALYST** Isacco Fusi **BUSINESS OPERATIONS ADMINISTRATOR** Taylor Fisher **DIGITAL/PRINT STRATEGY MANAGER** Jason Hillman **SENIOR MANAGER, PUBLISHING AND CONTENT SYSTEMS** Marcus Spiegler **ASSISTANT MANAGER DIGITAL/PRINT** Rebecca Doshi **SENIOR CONTENT & PUBLISHING SYSTEMS SPECIALIST** Jacob Hedrick **SENIOR CONTENT SPECIALIST** Lori Murphy **PRODUCTION SPECIALIST** Kristin Wowk **DIGITAL PRODUCTION MANAGER** Lisa Stanford **CONTENT SPECIALIST** Kimberley Oster **ADVERTISING PRODUCTION OPERATIONS MANAGER** Deborah Tompkins **DESIGNER, CUSTOM PUBLISHING** Jeremy Huntsinger **SR. TRAFFIC ASSOCIATE** Christine Hall **SPECIAL PROJECTS ASSOCIATE** Sarah Dhere

ASSOCIATE DIRECTOR, BUSINESS DEVELOPMENT Justin Sawyers **GLOBAL MARKETING MANAGER** Allison Pritchard **DIGITAL MARKETING MANAGER** Aimee Aponte **JOURNALS MARKETING MANAGER** Shawana Arnold **MARKETING ASSOCIATES** Aaron Helmbrecht, Ashley Hylton, Mike Romano, Tori Velasquez, Jenna Voris **SENIOR DESIGNER** Kim Huynh

DIRECTOR AND SENIOR EDITOR, CUSTOM PUBLISHING Sean Sanders **ASSISTANT EDITOR, CUSTOM PUBLISHING** Jackie Oberst **PROJECT MANAGER** Melissa Collins

DIRECTOR, PRODUCT & PUBLISHING DEVELOPMENT Chris Reid **DIRECTOR, BUSINESS STRATEGY AND PORTFOLIO MANAGEMENT** Sarah Whalen **ASSOCIATE DIRECTOR, PRODUCT MANAGMENT** Kris Bishop **PRODUCT DEVELOPMENT MANAGER** Scott Chernoff **PUBLISHING TECHNOLOGY MANAGER** Michael Di Natale **SR. PRODUCT ASSOCIATE** Robert Koepke **PRODUCT ASSOCIATE** Caroline Breul, Anne Mason **SPJ ASSOCIATE MANAGER** Samantha Bruno Fuller **SPJ ASSOCIATE** Casey Buchta

MARKETING MANAGER Kess Knight **BUSINESS DEVELOPMENT MANAGER** Rasmus Andersen **SENIOR INSTITUTIONAL LICENSING MANAGER** Ryan Rexroth **INSTITUTIONAL LICENSING MANAGER** Marco Castellan, Claudia Paulsen-Young **SENIOR MANAGER, INSTITUTIONAL LICENSING OPERATIONS** Judy Lillibridge **SENIOR OPERATIONS ANALYST** Lana Guz **SYSTEMS & OPERATIONS ANALYST** Ben Teincuff **FULFILLMENT ANALYST** Aminta Reyes

DIRECTOR, GLOBAL SALES Tracy Holmes **US EAST COAST AND MID WEST SALES** Stephanie O'Connor **US MID WEST, MID ATLANTIC AND SOUTH EAST SALES** Chris Hoag **US WEST COAST SALES** Lynne Stickrod **ASSOCIATE DIRECTOR, ROW** Roger Goncalves **SALES REP, ROW** Sarah Lelarge **SALES ADMIN ASSISTANT, ROW** Victoria Glasbey **DIRECTOR OF GLOBAL COLLABORATION AND ACADEMIC PUBLISHING RELATIONS, ASIA** Xiaoying Chu **ASSOCIATE DIRECTOR, INTERNATIONAL COLLABORATION** Grace Yao **SALES MANAGER** Danny Zhao **MARKETING MANAGER** Kilo Lan **ASCA CORPORATION, JAPAN** Yoshimi Toda (Tokyo), Miyuki Tani (Osaka)

DIRECTOR, COPYRIGHT, LICENSING AND SPECIAL PROJECTS Emilie David **RIGHTS AND PERMISSIONS ASSOCIATE** Elizabeth Sandler **LICENSING ASSOCIATE** Virginia Warren **CONTRACT SUPPORT SPECIALIST** Michael Wheeler

MAIN HEADQUARTERS
Science/AAAS
1200 New York Ave. NW
Washington, DC 20005

SCIENCE INTERNATIONAL
Clarendon House
Clarendon Road
Cambridge, CB2 8FH, UK

SCIENCE CHINA
Room 1004, Culture Square
No. 59 Zhongguancun St.
Haidian District, Beijing, 100872

SCIENCE JAPAN
ASCA Corporation
Sibaura TY Bldg. 4F, 1-14-5
Shibaura Minato-ku
Tokyo, 108-0073 Japan

EDITORIAL
science_editors@aaas.org

NEWS
science_news@aaas.org

INFORMATION FOR AUTHORS
science.org/authors/
science-information-authors

REPRINTS AND PERMISSIONS
science.org/help/
reprints-and-permissions

MEDIA CONTACTS
scipak@aaas.org

MULTIMEDIA CONTACTS
SciencePodcast@aaas.org
ScienceVideo@aaas.org

INSTITUTIONAL SALES AND SITE LICENSES
science.org/librarian

PRODUCT ADVERTISING & CUSTOM PUBLISHING
advertising.science.org/
products-services
science_advertising@aaas.org

CLASSIFIED ADVERTISING
advertising.science.org/
science-careers
advertise@sciencecareers.org

JOB POSTING CUSTOMER SERVICE
employers.sciencecareers.org
support@sciencecareers.org

MEMBERSHIP AND INDIVIDUAL SUBSCRIPTIONS
science.org/subscriptions

MEMBER BENEFITS
aaas.org/membership/benefits

AAAS BOARD OF DIRECTORS
CHAIR Claire M. Fraser
PRESIDENT Susan G. Amara
PRESIDENT-ELECT Gilda A. Barabino
TREASURER Carolyn N. Ainslie
CHIEF EXECUTIVE OFFICER Sudip Parikh
BOARD Cynthia M. Beall, Rosina M. Bierbaum, Ann Bostrom, Janine Austin Clayton, Laura H. Greene, Kaye Husbands Fealing, Maria M. Klawe, Robert B. Millard, William D. Provine

Science serves as a forum for discussion of important issues related to the advancement of science by publishing material on which a consensus has been reached as well as including the presentation of minority or conflicting points of view. Accordingly, all articles published in *Science*—including editorials, news and comment, and book reviews—are signed and reflect the individual views of the authors and not official points of view adopted by AAAS or the institutions with which the authors are affiliated.

BOARD OF REVIEWING EDITORS (Statistics board members indicated with **(S)**)

Erin Adams, *U. of Chicago*
Takuzo Aida, *U. of Tokyo*
Leslie Aiello, *Wenner-Gren Fdn.*
Deji Akinwande, *UT Austin*
Judith Allen, *U. of Manchester*
Marcella Alsan, *Harvard U.*
Sebastian Amigorena, *Inst. Curie*
James Analytis, *UC Berkeley*
Trevor Archer, *NIEHS, NIH*
Paola Arlotta, *Harvard U.*
David Awschalom, *U. of Chicago*
Delia Baldassarri, *NYU*
Nenad Ban, *ETH Zürich*
Nandita Basu, *U. of Waterloo*
Franz Bauer, *Pontificia U. Católica de Chile*
Ray H. Baughman, *UT Dallas*
Carlo Beenakker, *Leiden U.*
Yasmine Belkaid, *NIAID, NIH*
Philip Benfey, *Duke U.*
Kiros T. Berhane, *Columbia U.*
Joseph J. Berry, *NREL*
Alessandra Biffi, *Harvard Med.*
Chris Bowler, *École Normale Supérieure*
Ian Boyd, *U. of St. Andrews*
Malcolm Brenner, *Baylor Coll. of Med.*
Emily Brodsky, *UC Santa Cruz*
Ron Brookmeyer, *UCLA* **(S)**
Christian Büchel, *UKE Hamburg*
Dennis Burton, *Scripps Res.*
Carter Tribley Butts, *UC Irvine*
György Buzsáki, *NYU School of Med.*
Mariana Byndloss, *Vanderbilt U. Med. Ctr.*
Annmarie Carlton, *UC Irvine*
Simon Cauchemez, *Inst. Pasteur*
Ling-Ling Chen, *SIBCB, CAS*
M. Keith Chen, *UCLA*
Zhijian Chen, *UT Southwestern Med. Ctr.*
Ib Chorkendorff, *Denmark TU*
Amander Clark, *UCLA*
James J. Collins, *MIT*
Robert Cook-Deegan, *Arizona State U.*
Virginia Cornish *Columbia U.*
Carolyn Coyne, *Duke U.*
Roberta Croce, *VU Amsterdam*
Christina Curtis, *Stanford U.*
Ismaila Dabo, *Penn State U.*
Jeff L. Dangl, *UNC*
Chiara Daraio, *Caltech*
Nicolas Dauphas, *U. of Chicago*
Christian Davenport, *U. of Michigan*
Frans de Waal, *Emory U.*
Claude Desplan, *NYU*
Sandra Díaz, *U. Nacional de Córdoba*
Samuel Díaz-Muñoz, *UC Davis*
Ulrike Diebold, *TU Wien*
Stefanie Dimmeler, *Goethe-U. Frankfurt*
Hong Ding, *Inst. of Physics, CAS*
Dennis Discher, *UPenn*
Jennifer A. Doudna, *UC Berkeley*
Ruth Drdla-Schutting, *Med. U. Vienna*
Raissa M. D'Souza, *UC Davis*
Bruce Dunn, *UCLA*
William Dunphy, *Caltech*
Scott Edwards, *Harvard U.*
Todd Ehlers, *U. of Tübingen*
Nader Engheta, *UPenn*
Karen Ersche, *U. of Cambridge*
Beate Escher, *UFZ & U. of Tübingen*
Barry Everitt, *U. of Cambridge*
Vanessa Ezenwa, *U. of Georgia*
Toren Finkel, *U. of Pitt. Med. Ctr.*
Gwenn Flowers, *Simon Fraser U.*
Peter Fratzl, *Max Planck Inst. Potsdam*
Elaine Fuchs, *Rockefeller U.*
Jay Gallagher, *U. of Wisconsin*
Daniel Geschwind, *UCLA*
Ramon Gonzalez, *U. of South Florida*
Sandra González-Bailón, *UPenn*
Nicolas Gruber, *ETH Zürich*
Hua Guo, *U. of New Mexico*
Taekjip Ha, *Johns Hopkins U.*
Daniel Haber, *Mass. General Hos.*
Sharon Hammes-Schiffer, *Yale U.*
Wolf-Dietrich Hardt, *ETH Zürich*
Louise Harra, *U. Coll. London*
Carl-Philipp Heisenberg, *IST Austria*
Janet G. Hering, *Eawag*
Christoph Hess, *U. of Basel & U. of Cambridge*
Heather Hickman, *NIAID, NIH*
Hans Hilgenkamp, *U. of Twente*
Janneke Hille Ris Lambers, *ETH Zürich*
Kai-Uwe Hinrichs, *U. of Bremen*
Deirdre Hollingsworth, *U. of Oxford*
Randall Hulet, *Rice U.*
Auke Ijspeert, *EPFL*
Gwyneth Ingram, *ENS Lyon*
Darrell Irvine, *MIT*
Akiko Iwasaki, *Yale U.*
Stephen Jackson, *USGS & U. of Arizona*
Erich Jarvis, *Rockefeller U.*
Peter Jonas, *IST Austria*
Matt Kaeberlein, *U. of Wash.*
William Kaelin Jr., *Dana-Farber Cancer Inst.*
Daniel Kammen, *UC Berkeley*
Kisuk Kang, *Seoul Nat. U.*
Sabine Kastner, *Princeton U.*
V. Narry Kim, *Seoul Nat. U.*
Robert Kingston, *Harvard Med.*
Nancy Knowlton, *Smithsonian Institution*
Etienne Koechlin, *École Normale Supérieure*
Alex L. Kolodkin, *Johns Hopkins U.*
Julija Krupic, *U. of Cambridge*
Paul Kubes, *U. of Calgary*
Gabriel Lander, *Scripps Res.* **(S)**
Mitchell A. Lazar, *UPenn*
Luis Liz-Marzán, *CIC biomaGUNE*
Omar Lizardo, *UCLA*
Jonathan Losos, *Wash. U. in St. Louis*
Ke Lu, *Inst. of Metal Res., CAS*
Christian Lüscher, *U. of Geneva*
Jean Lynch-Stieglitz, *Georgia Inst. of Tech.*
David Lyons, *U. of Edinburgh*
Fabienne Mackay, *QIMR Berghofer*
Anne Magurran, *U. of St. Andrews*
Ari Pekka Mähönen, *U. of Helsinki*
Asifa Majid, *U. of York*
Oscar Marín, *King's Coll. London*
Charles Marshall, *UC Berkeley*
Christopher Marx, *U. of Idaho*
David Masopust, *U. of Minnesota*
Geraldine Masson, *CNRS*
C. Robertson McClung, *Dartmouth*
Rodrigo Medellín, *U. Nacional Autónoma de México*
C. Jessica Metcalf, *Princeton U.*
Baoxia Mi, *UC Berkeley*
Tom Misteli, *NCI, NIH*
Alison Motsinger-Reif, *NIEHS, NIH* **(S)**
Suresh Naidu, *Columbia U.*
Danielle Navarro, *U. of New South Wales*
Daniel Nettle, *Newcastle U.*
Daniel Neumark, *UC Berkeley*
Beatriz Noheda, *U. of Groningen*
Helga Nowotny, *Vienna Sci. & Tech. Fund*
Rachel O'Reilly, *U. of Birmingham*
Pilar Ossorio, *U. of Wisconsin*
Andrew Oswald, *U. of Warwick*
Isabella Pagano, *Istituto Nazionale di Astrofisica*
Elizabeth Levy Paluck, *Princeton U.*
Jane Parker, *Max Planck Inst. Cologne*
Giovanni Parmigiani, *Dana-Farber Cancer Inst.* **(S)**
Daniel Pauly, *U. of British Columbia*
Ana Pêgo, *U. do Porto*
Samuel Pfaff, *Salk Inst.*
Julie Pfeiffer, *UT Southwestern Med. Ctr.*
Philip Phillips, *UIUC*
Matthieu Piel, *Inst. Curie*
Kathrin Plath, *UCLA*
Martin Plenio, *Ulm U.*
Katherine Pollard, *UCSF*
Elvira Poloczanska, *Alfred-Wegener-Inst.*
Julia Pongratz, *Ludwig Maximilians U.*
Philippe Poulin, *CNRS*
Jonathan Pritchard, *Stanford U.*
Lei Stanley Qi, *Stanford U.*
Trevor Robbins, *U. of Cambridge*
Joeri Rogelj, *Imperial Coll. London*
Amy Rosenzweig, *Northwestern U.*
Mike Ryan, *UT Austin*
Miquel Salmeron, *Lawrence Berkeley Nat. Lab*
Nitin Samarth, *Penn State U.*
Erica Ollmann Saphire, *La Jolla Inst.*
Joachim Saur, *U. zu Köln*
Alexander Schier, *Harvard U.*
Wolfram Schlenker, *Columbia U.*
Susannah Scott, *UC Santa Barbara*
Anuj Shah, *U. of Chicago*
Vladimir Shalaev, *Purdue U.*
Jie Shan, *Cornell U.*
Beth Shapiro, *UC Santa Cruz*
Jay Shendure, *U. of Wash.*
Steve Sherwood, *U. of New South Wales*
Brian Shoichet, *UCSF*
Robert Siliciano, *JHU School of Med.*
Lucia Sivilotti, *U. Coll. London*
Richard Smith, *UNC* **(S)**
John Speakman, *U. of Aberdeen*
Tara Spires-Jones, *U. of Edinburgh*
Allan C. Spradling, *Carnegie Institution for Sci.*
V. S. Subrahmanian, *Northwestern U.*
Ira Tabas, *Columbia U.*
Eriko Takano, *U. of Manchester*
Patrick Tan, *Duke-NUS Med. School*
Sarah Teichmann, *Wellcome Sanger Inst.*
Rocio Titiunik, *Princeton U.*
Shubha Tole, *Tata Inst. of Fundamental Res.*
Maria-Elena Torres Padilla, *Helmholtz Zentrum München*
Kimani Toussaint, *Brown U.*
Barbara Treutlein, *ETH Zürich*
Jason Tylianakis, *U. of Canterbury*
Wim van der Putten, *Netherlands Inst. of Ecology*
Ivo Vankelecom, *KU Leuven*
Judith Varner, *UC San Diego*
Henrique Veiga-Fernandes, *Champalimaud Fdn.*
Reinhilde Veugelers, *KU Leuven*
Bert Vogelstein, *Johns Hopkins U.*
Julia Von Blume, *Yale School of Med.*
David Wallach, *Weizmann Inst.*
Jane-Ling Wang, *UC Davis* **(S)**
Jessica Ware, *Amer. Mus. of Natural Hist.*
David Waxman, *Fudan U.*
Chris Wikle, *U. of Missouri* **(S)**
Terrie Williams, *UC Santa Cruz*
Ian A. Wilson, *Scripps Res.* **(S)**
Hao Wu, *Harvard U.*
Li Wu, *Tsinghua U.*
Wei Xie, *Tsinghua U.*
Benjamin Youngblood, *St. Jude*
Yu Xie, *Princeton U.*
Jan Zaanen, *Leiden U.*
Kenneth Zaret, *UPenn School of Med.*
Bing Zhu, *Inst. of Biophysics, CAS*
Xiaowei Zhuang, *Harvard U.*
Maria Zuber, *MIT*

EDITORIAL

Africa's growing space enterprise

As the world becomes increasingly dependent on satellites and other space technology, new participants from every global region are joining the "space club" to reap the benefits—expanding communication and navigation capabilities; guiding policies with Earth observations; developing economies through growth in space ventures; and deepening our understanding of the cosmos through exploration. Consequently, countries are focusing on job opportunities, competitiveness, national security, and on having a voice in the global governance of space activities. Africa not only has a growing interest in space, but African countries are realizing that there is strength in pooling their resources and efforts to grow the continent's outer space enterprise.

In January 2016, the African Union (AU) adopted the African Space Policy and Strategy, with the goal of creating an integrated African space program that is responsive to the social, economic, political, and environmental needs of the continent. Another goal is to develop a regulatory framework that supports Africa's competitiveness in space while ensuring that Africa is a responsible and peaceful user of outer space. Two years later, the African Space Agency was established to implement continental space policy and strategy, as well as to oversee the support and coordination of space programs across Africa. This includes facilitating access to space resources and services, developing space infrastructure, and strengthening space research and development. Such aims require local and international cooperation.

> "...past international efforts...tended to cast Africa as a resource for external interests..."

The African Space Agency does face challenges. The agency was officially established as an organ of the AU Commission, rather than as a specialized agency with political and financial autonomy, which limits its opportunities to build international partnerships and fund projects. Another challenge is how the agency will operate with national space agencies across the continent. Fourteen African countries now have established national space agencies, and 9 countries have national space policy frameworks. Such independent efforts are expected to increase over the next few years. The African Space Agency needs to create a framework for leveraging the capabilities of each country without conflicting with their national obligations.

The good news is that intra-African cooperation is strong and growing. For example, Smart Africa is an alliance of 30 African countries that aims to ramp up satellite-based internet connectivity. More public-private partnerships across Africa intend to bridge equity gaps by providing connectivity and enabling access to the space market. Another example is the Square Kilometer Array (SKA), the world's largest and most advanced radio telescope, currently under development. Nine African countries are collaborating to carry out this project (Botswana, Ghana, Kenya, Madagascar, Mauritius, Mozambique, Namibia, South Africa, and Zambia). Fourteen countries from around the world, including South Africa, constitute The SKA Organisation, which will supply funds and technical expertise for the facility. Data from the SKA will be used by scientists around the world.

With a growing space enterprise, Africa must develop an indigenous workforce to stay competitive. Already, 13 African countries have launched 44 satellites at a cost of $4.5 billion, with 125 more in development. The AU needs to accelerate launching the Pan-African University Institute for Space Sciences, which is intended to fill a pipeline with skilled labor for Africa's space industry. Given the current impasse between the AU and South Africa, which was given hosting rights, the AU should begin a new process to choose another host country and quickly establish this much needed institution. Furthermore, an African space education summit should be convened to build a strategy to prepare the next generation of students to enter all sectors of the space enterprise. The African Academy of Science, with its expertise in geological, environmental, Earth, and space science, as well as policy, is in a good position to start planning that meeting.

A vibrant space industry is a refreshing change from past international efforts that tended to cast Africa as a resource for external interests or simply as a continent always in need of support. US Secretary of State Antony Blinken emphasized that Africa has a choice of countries to partner with, and should expect better engagement, hinting at the need for broad partnerships and investments in Africa. The question of what this could look like should be put on the agenda at the next US-Africa Summit later this year. African nations are now in a position to deliver the benefits of space innovation to their populations and to the world.

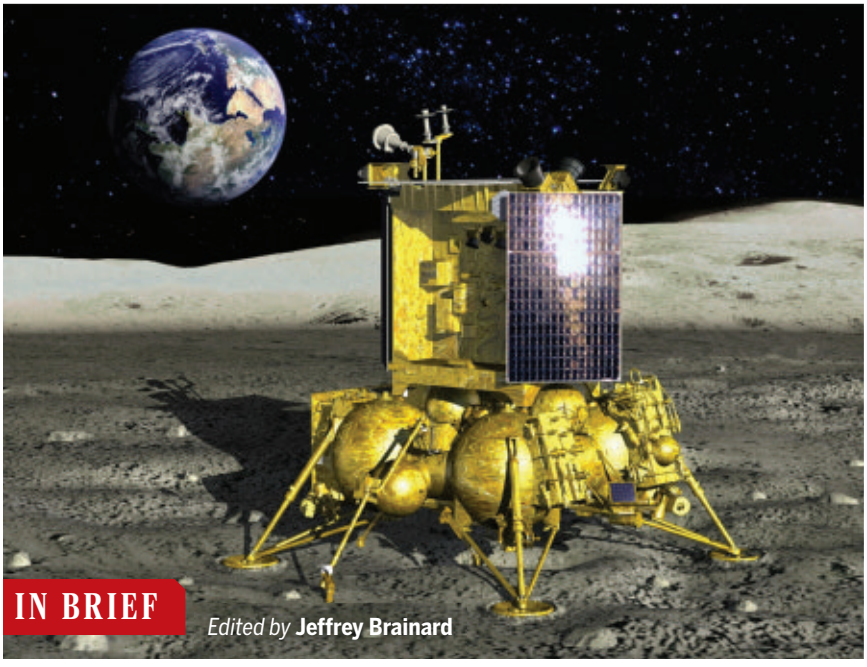
–Timiebi Aganaba and Etim Offiong

Timiebi Aganaba is an assistant professor at the School for the Future of Innovation in Society, Arizona State University, Tempe, AZ, USA. taganabajeanty@asu.edu

Etim Offiong is an assistant chief scientific officer at the African Regional Center for Space Science and Technology Education in English, Ile-Ife, Nigeria. eooffiong@arcsstee.org.ng

10.1126/science.abq5570

> “There should be a mandatory course in public health in our nation’s law schools.”
>
> **Vaccine researcher Peter Hotez** about a U.S. district court ruling that struck down the Centers for Disease Control and Prevention’s COVID-19 mask mandate on public transportation.



IN BRIEF

Edited by **Jeffrey Brainard**

The European Space Agency asked Russia to remove a camera experiment from its Luna-25 Moon lander.

LUNAR SCIENCE

Europe quits Russian Moon missions

The European Space Agency (ESA) announced last week it will stop partnerships with Russia on three upcoming Moon missions because of the war in Ukraine. The agency said it would shift instruments scheduled to fly on Russia’s Luna-25, -26, and -27 probes to lunar missions planned by Japan and the United States. ESA asked Roscosmos, the Russian space agency, to remove an instrument—a camera called Pilot D—that was already installed on the Luna-25 robotic lander, which is due for launch later this year. The decision came after ESA announced in March it was suspending work with Russia on the ExoMars rover mission.

EPA ups formaldehyde warning

TOXICOLOGY | A long-delayed assessment of the health effects of formaldehyde has concluded that the widely used chemical poses a greater cancer risk than had been estimated, which could lead to greater regulation. Most people are exposed to formaldehyde in airborne emissions from the glue in treated plywood and particle board used in buildings. A draft assessment 12 years ago by the U.S. Environmental Protection Agency (EPA) that linked formaldehyde exposure to an increased risk of leukemia and other cancers drew objections from members of Congress and industry representatives, who maintain that products containing formaldehyde are safe. The new draft, released last week, finds a higher risk of myeloid leukemia. In an unusual step, it will be reviewed by the National Academies of Sciences, Engineering, and Medicine, which 11 years ago criticized EPA for not clearly presenting its methods in the earlier assessment. If the agency formally classifies formaldehyde as a carcinogen, EPA could tighten restrictions on its use under the Toxic Substances Control Act, although the regulatory process is slow.

Sub-subvariants hit New York state

COVID-19 | Two new lineages of SARS-CoV-2, descendants of an already rapidly spreading spinoff of the Omicron variant of concern, have surged in central New York state, accounting for more than 90% of infections there by 13 April—the first reported instance of significant community spread for the lineages in the United States, authorities said. The New York State Department of Health said last week that BA.2.12 and BA.2.12.1 are more contagious than their parent, the Omicron subvariant BA.2, and are “likely contributing to the rising cases.” As of 1 April, infection and hospitalization rates in central New York were more than double the state average. No evidence so far suggests these latest versions of Omicron make people sicker, the state said. BA.2.12.1 has a spike protein mutation, L452Q, that increased the infectivity of a previous variant of interest, Lambda. BA.2.12.1 was first identified in mid-January in Canada, where 63 cases have now been found, and a few cases have been detected in five other countries besides the United States, but only in single-digit numbers. More than 1900 U.S. cases have been identified.

U.K. tests antibiotic subscriptions

DRUG DEVELOPMENT | The United Kingdom is moving closer to a world first: paying pharma companies a subscription-style fee for access to new antibiotics. The model is designed to give companies a financial incentive to innovate. New antibiotics are needed because their widespread use in human and veterinary medicine has led to the spread of resistant bacteria. But pharma companies see little financial reward in developing novel antibiotics, which are usually held in reserve for cases in which other antibiotics do not work, and then given for

ILLUSTRATION: N.P.O. LAVOCHKIN

Volunteers deposit recycled oyster shells for a reef restoration project in Virginia's Piankatank River.

PHOTO: AILEEN DEVLIN/VIRGINIA SEA GRANT

MARINE ECOLOGY

Rebuilt oyster reefs thrive, benefiting ecosystems

Oyster reefs can be restored and gain important ecological functions in coastal zones in as few as 6 years and persist, according to a long-term study. Oysters filter water, and reefs made of their shells provide habitats for young fish and other creatures. Beginning in the 19th century, many U.S. reefs were overharvested for food, and diseases wiped out others. After 2000, publicly and privately funded groups expanded projects to rebuild reefs by seeding them with oyster shells and larvae—but scientists have typically not monitored results for more than a year or two. A research team at the University of Virginia tracked dozens of reefs in coastal Virginia created by the Nature Conservancy during the past 15 years, in work supported by the National Science Foundation's Long-Term Ecological Research program. Many have become just as large and densely populated as nearby, surviving natural reefs, the team reported this month in *Conservation Letters*. The rebuilt reefs are stable in storms, and mud crabs—which serve as food for large fish—are now just as abundant as elsewhere.

just a few days. The subscription model provides guaranteed income for drugs that have been proven efficacious and licensed. A test of the subscription model could begin soon using two new antibiotics, one developed by Japanese drugmaker Shionogi and the other by U.S. manufacturer Pfizer. In an important milestone, the U.K. National Institute for Health and Care Excellence concluded last week that the drugs' medical value was worth paying each company £10 million yearly for up to 10 years. Negotiations with the companies are underway.

Team finds Notre Dame relics

ARCHAEOLOGY | Scientists have discovered a trove of significant artifacts and tombs beneath the floor of Notre Dame de Paris Cathedral, some of which may date to the 13th century. The excavation was to help prepare a section of the floor for scaffolding to support the planned replacement of the iconic cathedral's spire, which toppled in a catastrophic 2019 fire. The team found 10 sarcophagi, which it estimates were buried between the 14th and 18th centuries, the French National Institute of Preventive Archaeological Research said last week in a summary. Scientists peeked through a crack in one of the vessels using a tiny camera and saw a skeleton and what appeared to be fabric and hair. Scientists plan to conduct DNA and carbon-dating tests. The team also recovered well-preserved painted sculptures and many pieces of the medieval rood screen, which once divided congregants from clergy until it was demolished under Louis XIV.

Young kids get mystery hepatitis

INFECTIOUS DISEASES | Scientists are investigating an outbreak of severe hepatitis that has sickened dozens of formerly healthy young children in Western Europe and the U.S. state of Alabama. The viruses that commonly cause hepatitis were not found in the children, but many were carrying adenovirus, which usually causes colds and conjunctivitis. None was vaccinated against COVID-19, and several had the disease at or shortly before hospital admission. At least nine children have required liver transplants.

NSF seeks safer field studies

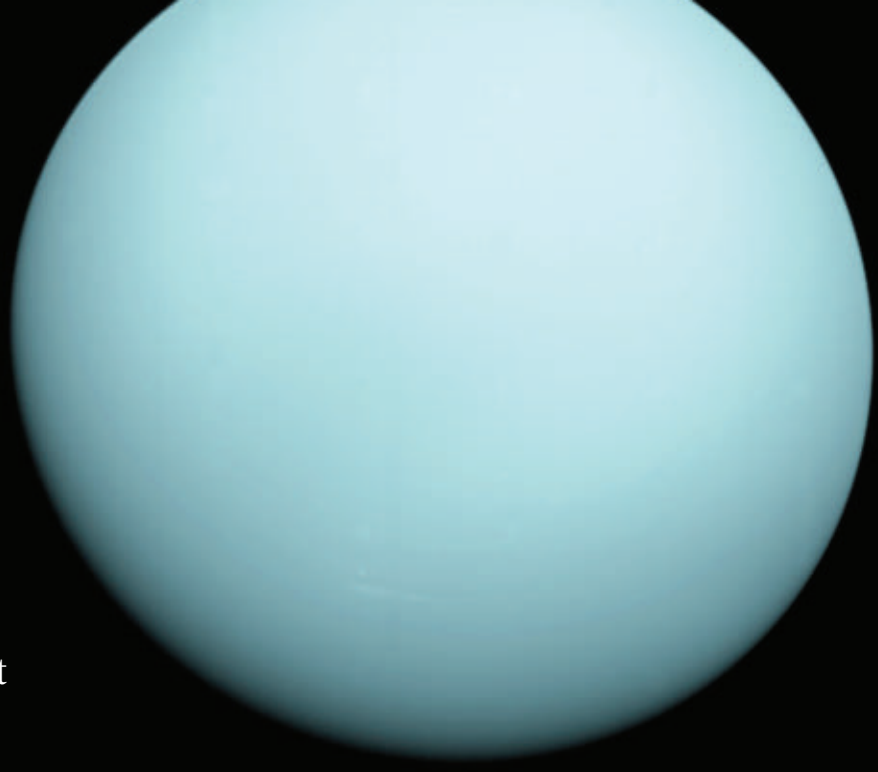
#METOO | The U.S. National Science Foundation (NSF) pledged last week to better police sexual and other forms of harassment at field sites and on research vessels, as part of a push on diversity by President Joe Biden's administration. In an "equity action plan" released on 14 April, the $8.8 billion agency noted that field sites and vessels have been associated with higher risks of harassment, especially for minorities within a field team. It said it would "engage in [harassment] monitoring, evaluation, and mitigation activities" with its grantees. The agency, a major funder of field studies, published on 13 April proposed changes to its grant awards procedures, requiring applicants to submit a written plan for maintaining safe, inclusive fieldwork sites.

IN DEPTH

PLANETARY SCIENCE

Uranus should be NASA's top target, report finds

Influential survey recommends $4.2 billion flagship mission to the ice giant

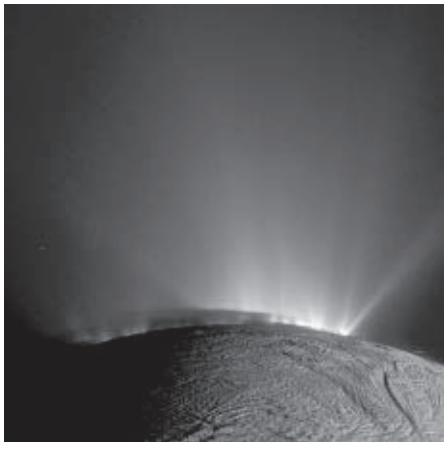


By **Paul Voosen**

After decades in the shadow of the other planets, Uranus should become NASA's focus of exploration, a panel of planetary scientists reported this week in the field's long-awaited "decadal survey," a priority-setting report the agency will use to makes its case to congressional funders. If the scientists get their wishes, NASA in the early 2030s will launch a $4.2 billion orbiter and atmospheric probe to Uranus, seeking to understand the formation and composition of this ice giant. Intermediate in size between the rocky planets and gas giants, Uranus and its neighbor Neptune "represent a unique planetary type that we poorly understand," says Ravit Helled, a planetary scientist at the University of Zürich, one of 130 scientists who contributed to the survey.

The decision to favor Uranus over Neptune came down to celestial opportunism, says Robin Canup, a planetary scientist at the Southwest Research Institute and co-chair of the report, which was overseen by the National Academies of Sciences, Engineering, and Medicine. If launched on a Falcon Heavy rocket in 2031 or 2032, the orbiter could get a gravity assist from Jupiter and arrive in 13 years; Neptune would take far longer. "This mission is technically ready to go," Canup says. "We advocate that it be started right away." But whether that can happen depends on NASA's ability to manage a budget that has been strained by the pandemic and soaring mission costs.

It was Uranus's turn. The last decadal report, in 2011, ranked an ice giants mission third, following a set of missions to return rock samples from Mars and a visit to Europa, Jupiter's icy moon—projects that are now underway or in development. So perhaps the survey's biggest surprise is its recommendation for what comes after Uranus: a $4.9 billion mission to Enceladus, a tiny moon of Saturn that spews organic-rich plumes of water out of fissures in an icy cap—ready-made samples of a subsurface ocean that might host microbes. "Enceladus is probably the best place to look for evidence of life that we can do today," says Philip Christensen, a planetary scientist at Arizona State University, Tempe, and the report's other co-chair.



A mission to Enceladus—a moon of Saturn that spews saltwater into space—was ranked No. 2 by a survey.

Uranus has not been explored up close since a flyby during the Voyager mission in 1986.

(The recommendation will mark an end for plans to put a lander on Europa's surface, which had previously been advanced as a top future mission.)

The report also lists targets for a set of large missions that scientists can propose to NASA, called New Frontiers. Some concepts are familiar from past surveys: a Saturn probe, a comet sample return, a lunar geophysical network. Others are new: sample return from Ceres, the water-rich dwarf planet in the asteroid belt; an orbiter and lander to a Centaur, one of the small bodies between Jupiter and Neptune believed to capture the composition of the early Solar System; an orbiter of Saturn's moon Titan; a Venus lander; and an Enceladus plume sampler. (Enceladus's inclusion in two different mission categories stresses its importance, Christensen says.)

NASA should also continue programs dedicated to exploring the Moon and Mars, the panel recommends. After the agency builds the Mars sample return missions, the panel calls for it to develop a $1.1 billion robotic lander, called the Mars Life Explorer, that would drill 2 meters into midlatitude ice deposits.

For the Moon, the panel endorses the Artemis program, funded by NASA's human spaceflight division, which plans to return astronauts to the surface. But it suggests science should drive the choices of what to do, rather than being an afterthought. "It's not just flags

PHOTOS: (TOP TO BOTTOM) NASA/JPL; NASA/JPL-CALTECH/SPACE SCIENCE INSTITUTE

and footprints," says Bethany Ehlmann, a planetary scientist at the California Institute of Technology and co-author of the report. The report calls for a $1.5 billion long-range large robotic rover called Endurance-A that could cover 1000 kilometers, drill 100 kilograms worth of samples, and return them to astronauts who would eventually bring them back to labs on Earth.

Those ambitions will strain NASA's planetary science budget, now $3.1 billion per year—the highest since the Viking missions to Mars in the 1970s. The Mars sample return campaign, which will retrieve rocks collected by the Perseverance rover, will cost more than $7 billion and consume one-fourth of the planetary budget in the next few years. The cost of Europa Clipper, which after launch in 2024 will swoop past the moon nearly 50 times, has grown from $4.25 billion to $5 billion. And several cost-capped competitive missions have seen their budgets more than double because of the time needed to reach their remote destinations, a factor not included in their spending limits.

The budget overruns have led NASA to postpone missions: The ambitious Dragonfly rotocopter to Titan, Saturn's methane-rich moon, will now launch in 2027 instead of 2025, and the next New Frontiers selection will be delayed by several years. To stop this cycle of overruns, the report suggests, NASA needs to face reality and raise the cost caps for the two competitive mission lines, New Frontiers and Discovery, to $1.65 billion and $800 million, respectively, while also forcing those missions to fully account for lifetime costs. Those measures should still allow NASA to select five Discovery missions over a decade, but only one New Frontiers mission.

Although the budget has grown to accommodate big missions, the scientists who advance that work have not seen the same gains, the report stresses. The share of the budget spent on research grants has fallen from 14% in 2010 to 7.7%. Progress has been made in recruiting more women to the field, but less so with underrepresented racial and ethnic groups. Latino and Black scientists make up just 5% and 1% of the planetary science workforce. "We have untapped talent and we're missing out on great people and great ideas," Canup says. The report recommends collecting better demographic data and expanding predoctoral programs that support students from underrepresented communities.

Students entering the field now could constitute the scientific heart of the mission targeting Uranus, which humanity first saw up close with the Voyager 2 flyby in 1986. That survey prompted many scientists to think of the ice giants as anomalies: stunted gas giants that accumulated only a couple Earth masses' worth of hydrogen and helium during formation, either because of a lack of gas or a late start. But since Voyager, astronomers have found thousands of planets around others stars, and many are Uranus-size, says Jonathan Fortney, a planetary scientist at the University of California, Santa Cruz. "Nature loves to make planets of this size."

Uranus also holds its own individual appeal. Its spin axis is nearly horizontal—likely the result of a giant impact early in its history that tipped it over. Compared with the other planets, it gives off little heat, suggesting either that it is surprisingly cold or that its atmosphere has put a lid on escaping heat. It has two sets of rings, along with a densely packed set of primordial moons and oddball objects, likely trapped comets or objects from a region beyond Neptune called the Kuiper belt. "Some may still have water on the inside," says Kirby Runyon, a planetary scientist at the Johns Hopkins University Applied Physics Laboratory (APL).

"Enceladus is probably the best place to look for evidence of life that we can do today."
Philip Christensen, Arizona State University, Tempe

For the lure of an ocean, however, it's hard to top tiny Enceladus, just 504 kilometers wide. In 2005, NASA's Cassini spacecraft spotted plumes of saltwater erupting from rifts in its icy surface. Subsequent flights through those plumes revealed abundant organic molecules, necessary to build life, along with silica and hydrogen gas, a sign that the ocean feeding the plumes probably has hydrothermal vents in its depths, a potential energy source for microbes (*Science*, 14 April 2017, p. 121).

The survey endorsed a hybrid "orbilander" mission to Enceladus, which would sample the plume and survey the moon's surface for a couple years before turning on its side and landing, a relatively easy task in a place with weak gravity and no appreciable atmosphere. It would target a place where the erupting water falls as snow, which its instruments could sample. Two would explicitly be aimed at detecting life: a DNA sequencer and a microscope. Enceladus has checked off all the requirements for habitability, says Shannon MacKenzie, a planetary scientist at APL who led a study developing the idea. "The next question is: Is Enceladus inhabited?" ■

COVID-19

China refuses to end harsh lockdowns

Still not fully prepared for living with the virus, the government is sticking to "zero COVID"

By **Dennis Normile**

A haunting video that went viral this month showed residents of Shanghai screaming from high-rise windows into the night—a collective complaint about the harsh COVID-19 lockdown the city's 26 million inhabitants have been under since the end of March. Many have had trouble obtaining food or essential medicines. Tens of thousands who have tested positive for SARS-CoV-2 are isolated in crowded halls and convention centers outfitted with cots in cubicles, without showers. Authorities have separated infected children from their parents. Media have reported deaths among patients denied health care for conditions other than COVID-19.

But the Chinese government is not budging. Although the country's public health experts have worked quietly on steps toward coexisting with SARS-CoV-2 (*Science*, 3 March, p. 949), as much of the world has begun to do, President Xi Jinping reiterated on 13 April that the country must persist with its "zero COVID" strategy, now called "dynamic clearing," despite the growing costs.

That insistence reflects a fear of an explosion of serious illness and death if Omicron variants of SARS-CoV-2, now largely one called BA.2, escape control. More than 2 years into the pandemic, China is still not fully prepared; its leadership has squandered the grace period it earned with the zero COVID strategy, says Yanzhong Huang, a global health specialist at the Council on Foreign Relations, a U.S. think tank. "Until November last year, there was no serious effort to prioritize the vaccination of the elderly," he says. Nor did China use the time to improve the rural health infrastructure.

The risks aside, a reversal is politically sensitive for the Chinese government, which has long boasted about its superior strategy. By now, the COVID-19 response has become "not so much a public health or public policy issue as a political issue," Huang says. "It has be-

Workers converted Shanghai's National Exhibition and Convention Center into a COVID-19 hospital in early April.

come an undebatable political decision," adds Xi Chen, a public health scientist at the Yale School of Public Health.

But whether China can still get back to zero, as it did after the world's first COVID-19 outbreak in Wuhan in 2020 and every resurgence elsewhere since, is uncertain given Omicron's transmissibility. On 18 April, the daily number of reported cases in Shanghai had dropped somewhat, to 20,416, from a peak of 28,145 cases on 13 April. But the lockdown has been extended indefinitely. Meanwhile, other cities have entered partial or complete shutdowns as well.

As elsewhere, Shanghai's Omicron epidemic appears to be mild. Roughly 90% of cases are reportedly asymptomatic, and by 19 April, the city had only reported 10 deaths. (In all of China, only 12 COVID-19–related deaths have been reported this year.)

Whether the numbers tell an accurate story is unclear. One reason for the large number of asymptomatic cases is the repeated testing of Shanghai residents, which catches infections that would fly under the radar in other countries. But Huang suspects Shanghai officials are also counting people who only have coldlike symptoms as asymptomatic cases, because they can be sent to the makeshift isolation centers instead of hospitals.

As to severe cases and deaths, local media have raised questions about unexplained deaths at several Shanghai nursing homes. Still, Chen believes the numbers are "largely reliable." He says severity is low because most infections so far seem to have occurred in younger people and because the outbreak is still in its early phase. Also, identifying infections early allows for timely treatment that can forestall serious illness. "If infections spill over to sizable older populations, the results will be very different from what we are seeing now," and more similar to Hong Kong, Chen says. That city recorded the world's highest death rate in mid-March, in part because many older people had foregone the vaccine.

Mainland China has a similar problem. Airfinity, a London-based health analytics firm, reported that nearly 19% of Chinese people over age 60 were unvaccinated as of mid-March. Among those in their 80s, just over 50% had received two shots by 18 March, and only 19% had received boosters, Zeng Yixin, deputy director of the National Health Commission, reported at a press briefing. (Across all ages, the picture looks better, with 88% of the population fully vaccinated and about 51% having received boosters.)

If Omicron spreads throughout China, the low vaccination rate among the elderly could lead to 1 million deaths in 3 months, Airfinity estimates. In contrast, Australia, New Zealand, and Singapore exited zero COVID after only 1.2%, 0.6%, and 0.4%, respectively, of their over-60 population were unvaccinated. They have seen relatively few deaths as a result.

Why China did worse is puzzling. Some say the country's leadership painted itself into a corner by touting the success of zero COVID and not preparing the public for a transition to living with the virus. Because of safety concerns about COVID-19 vaccines, the campaign initially focused on younger populations. Vaccine hesitancy continues even today. "One question I frequently get asked is: 'My parents are 80 years old [or] have hypertension-diabetes, is it safer for them to vaccinate or not vaccinate?'" says Zhangkai Cheng, a respiratory specialist at Guangzhou Medical University.

"Hesitancy among older adults is somehow rational in the context of a zero COVID strategy," says University of Hong Kong (HKU) epidemiologist Ben Cowling. "If a vaccine has minimal but nonzero risk, whereas the risk of COVID infection might be expected to be minimal, some may prefer to take their chances with the virus." Now that this calculation has changed, some communities are sending vaccination teams door-to-door to persuade senior citizens to get the shots, offering gift cards as an incentive.

A study by an HKU group provided some reassurance about the effectiveness of the inactivated virus vaccines used in China, which has not authorized the messenger RNA (mRNA) shots used in many other countries. The effectiveness of two doses of Sinovac's inactivated vaccine was notably lower than that of two doses of an mRNA vaccine among adults 60 and older, the team found, but three doses of either vaccine offered very good protection against severe illness and death.

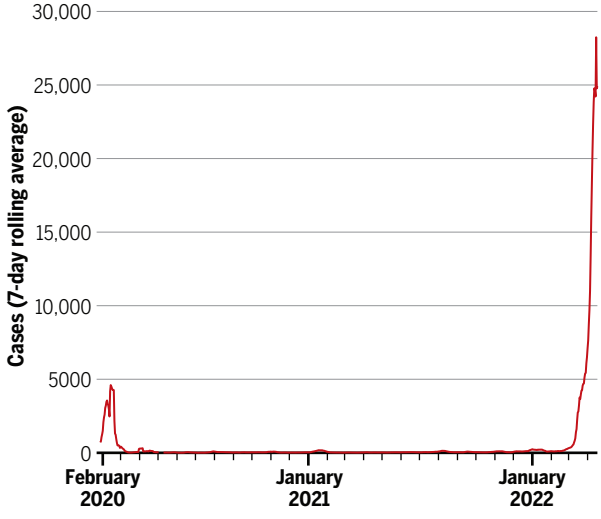
China's plans for exiting its zero COVID stance are still in development. In a letter published in *Nature Medicine* on 5 April, two Chinese epidemiologists listed several measures the country needs to implement to make the shift this year, including expanding the use of online health consultations, training health care workers to better treat mild COVID-19 cases, and stockpiling antivirals.

But few are willing to wager when the transition will occur. "There is no indication that the central government has begun to prepare for a coexistence strategy," says Xi Lu, a National University of Singapore specialist in Chinese economic policy. "I speculate that China will continue to implement the wrong policies for a long time; and with each day of delay, the transition will become more difficult." ■

With reporting by Bian Huihui.

To zero and back

China's "zero COVID" strategy has been successful for 2 years, but now cases of the highly transmissible Omicron variant are surging.



CREDITS: (PHOTO) DING TING/XINHUA VIA GETTY IMAGES; (GRAPHIC) K. FRANKLIN/*SCIENCE*; (DATA) OUR WORLD IN DATA

PALEONTOLOGY

Pterosaurs were clad in colorful plumage

Study suggests feathers arose—and were used for display—well before reign of dinosaurs

By **Rodrigo Pérez Ortega**

The idea that dinosaurs sported colorful feathers, once outlandish, has become conventional wisdom. Now, a new study of a Brazilian fossil suggests that pterosaurs—leathery winged, flying reptiles only distantly related to dinosaurs—were also clad in tiny feathers of varying hues. The finding suggests feathers may have evolved more than 150 million years before the heyday of the dinosaurs, probably for display, the authors say. "In their very earliest forms, feathers were colored … presumably for signaling," says paleobiologist Maria McNamara of University College Cork, who led the study.

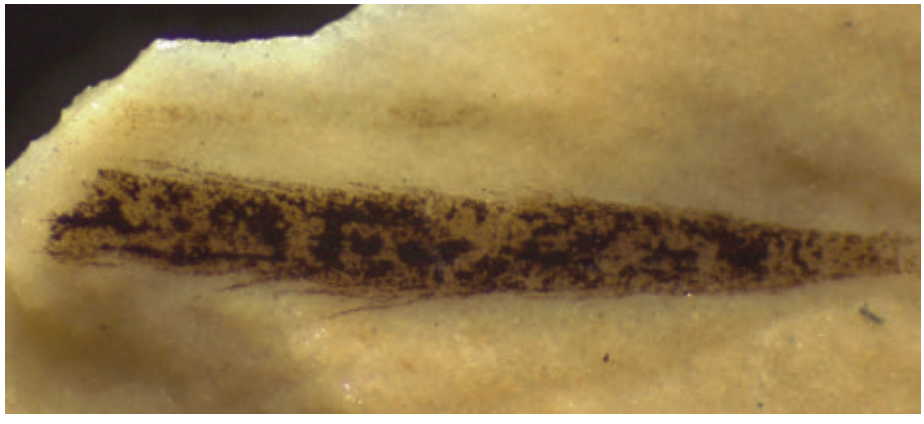
The paper "reinforces the idea that pterosaurs were 'fluffy,' and indicates at least some of them probably had complex colorful patterns—which is fantastic," says Rodrigo Pêgas, a paleontologist at the Federal University of ABC, São Bernardo do Campo, in Brazil. But Pêgas is not convinced that feathers originated as early as McNamara thinks—and some other researchers doubt the structures are feathers at all.

How feathers arose has been a big question in paleontology for more than 150 years, since the first *Archaeopteryx*—a feathered dinosaur once thought to be the first bird—was found in Germany. Many researchers think feathers arose for insulation and were co-opted only much later for flight and other uses, such as courtship displays. As for pterosaurs, researchers had previously reported their bodies were covered in pycnofibers, single-stranded structures that formed a "fuzz," presumably for warmth.

Then in 2018, McNamara and her colleagues reported that two well-preserved Chinese pterosaurs showed what seemed to be a defining feature of feathers: a central shaft with branches. Some paleontologists were skeptical, and McNamara says she understood why. "Their feathers were—to be honest—a bit weird," she says. "They didn't branch like modern bird feathers do."

Now, she and her colleagues have cemented their arguments with a paper this week in *Nature* analyzing the soft tissue of an exquisitely preserved skull of *Tupandactylus imperator*—a pterosaur that had a majestic head crest and a 5-meter wingspan. It lived 113 million years ago in what is now the Araripe Basin in northeastern Brazil, although McNamara studied the fossil in Belgium. The team thinks it was poached from Brazil and kept in private collections until recently. Earlier this year, the Royal Belgian Institute of Natural Sciences repatriated the fossil to Brazil, where it will be displayed at the Earth Sciences Museum in Rio de Janeiro. "It is great that the fossil is back in Brazil," Pêgas says.

On the pterosaur's head crest, the researchers identified both single-stranded fibers and featherlike branching ones with a central shaft narrowing at the base. Under the scanning electron microscope, both skin and feathers had melanosomes, intracellular structures containing melanin that give pigment to skin, feathers, and fur in living animals, with differently shaped melanosomes conferring different colors. The pterosaur's melanosomes had diverse shapes—ovid, spherical, and elongated—something until now only seen in mammalian fur and dinosaur and bird feathers.



This tiny feather, about 1.5 millimeters long, adorned the head crest of a 113-million-year-old pterosaur.

PHOTO: AUDE CINCOTTA

The researchers think *Tupandactylus*'s colored, branching structures were indeed feathers, which both kept it warm and enabled it to signal to other pterosaurs, perhaps as male peacocks do by displaying plumage during mating.

The finding means feathers must have evolved far earlier than was thought, McNamara says. "The most parsimonious explanation is that feathers were present in the common ancestor of [pterosaurs and dinosaurs]," about 250 million years ago during the Triassic period.

Some paleontologists say the evidence of feathers is persuasive. "We're hammering it in with 7-inch nails with these findings," says Jakob Vinther, a paleobiologist at the University of Bristol. Paleontologist Michael Benton, also at Bristol, agrees, but "I don't think pterosaur feathers had any function in flight because they're just fluffy little feathers."

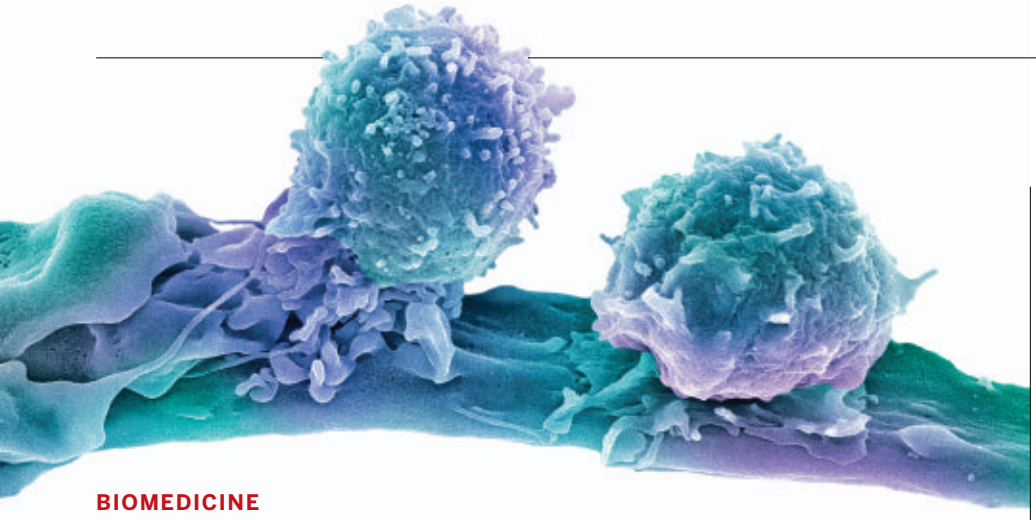
But paleontologist David Martill from the University of Portsmouth says the small branched structures "look nothing like feathers." He thinks they are a different kind of keratinous covering, though he agrees they were probably spectacularly colored.

Even if they are feathers, pterosaurs may not have flaunted them like Mesozoic peacocks, Vinther says. He notes that the researchers didn't infer the melanosomes' color and says it's possible the plumage was used for camouflage rather than display.

Nor is it certain that the pterosaur structures share an ancient origin with those of dinosaurs and their descendants, living birds, some researchers say. "We still need fossil evidence for feathers in the Triassic as well as unequivocal molecular evidence for the common origin between pterosaur pycnofibers and dinosaur feathers," Pêgas says.

McNamara promises more evidence for her scenario. Her team is working to characterize the detailed chemistry of the *Tupandactylus* samples, which could reveal organic compounds in the feathers.

If the current findings hold up, they may shed light on the selection pressures that shaped early feathers, says Jasmina Wiemann, a molecular paleobiologist at the California Institute of Technology. "Thermal regulation has been the old hypothesis out there … [but] maybe there's more to it." ■



T cells (spheres) may attack a tumor better if the cancer has deliberately added mutations.

BIOMEDICINE

Fight mutant cancer cells by … making more mutations?

Chemotherapies that induce new DNA errors may help unleash the immune system on tumors

By **Jocelyn Kaiser**, *in New Orleans*

Researchers are devising ways to paint targets on cancer cells. Drugs that unleash the immune system against cancer can be powerfully effective, but they appear to work best on the subset of tumors that are most riddled with mutations. Enter a controversial solution: Use chemotherapy to deliberately create new mutations in tumors and thereby make them more vulnerable to an immune system attack.

Lab studies and several small clinical trials already hint the strategy may help. "There might be an opportunity to begin to remodel the genetics of the tumor in such a way" that immunotherapy works better, a leader of one trial, cancer geneticist Luis Diaz of Memorial Sloan Kettering Cancer Center, said at a plenary session here at the annual meeting of the American Association for Cancer Research (AACR).

Still, some cancer researchers are leery of purposely inducing mutations and say animal experiments suggest doing so could cause more harm than good. "I question the rationale," says melanoma immunotherapy researcher Antoni Ribas of the University of California, Los Angeles.

Drugs called checkpoint inhibitors remove a molecular brake that keeps immune sentries called T cells from attacking tumors. They work best on cancers such as lung tumors triggered by smoking-induced DNA damage and melanomas, which accumulate mutations from ultraviolet (UV) light. Many of these genetic changes cause cells to make "neoantigens," novel protein fragments on tumor cells that flag them to T cells.

The notion that forcing cancer cells to make more neoantigens might bolster immunotherapy traces back to studies of tumors with defects in certain mechanisms that repair DNA. These cancer cells accumulate many mutations, and in 2015 a team led by Diaz, then at Johns Hopkins University, reported that checkpoint drugs work well on multiple tumor types with these "mismatch" DNA repair defects.

Cancer geneticist Alberto Bardelli of the University of Turino and colleagues went further by deliberately inactivating a mismatch repair gene in tumor-bearing mice. They reported in *Nature* in 2017 that the change resulted in a buildup of DNA errors in the cancer cells and boosted the effectiveness of checkpoint drugs.

Since then, two Italian trials have documented similar effects in people. One study gave the standard chemotherapy drug temozolomide, which disables mismatch repair genes, to 33 people with advanced colon cancer, which normally does not respond to checkpoint inhibitors because it has too few mutations. The chemotherapy alone shrank tumors in eight people but another seven people similarly responded after all later received two checkpoint inhibitors. Tumor growth halted in the overall group for an average of 7 months, the team reported last month in the *Journal of Clinical Oncology*.

In four patients who had tumor biopsies analyzed, as well as in 14 of 16 patients in a trial described in a poster at the AACR meeting, the team showed that temozolomide had induced mutations. Bardelli says the preliminary data offer a "proof of concept."

Diaz wondered whether inducing a specific kind of mutation would work even better. His team was particularly interested in a type that shifts how a cell's proteinmaking machinery reads a gene's messenger RNA. Such a "frameshift" mutation can change many of the amino acids of a gene's protein, making it more foreign to the immune system.

Postdoc Benoit Rousseau and others in the Diaz lab tested temozolomide and another chemotherapy drug, cisplatin, on cancer cells and found the combination produced 1000 times more frameshift mutations than either drug alone. When cancer cells treated with the drug combination were injected into mice, the resulting tumors vanished in response to a checkpoint drug.

Diaz's team is now giving the combination to people with metastatic colon tumors before they receive a checkpoint drug. In two of the first 10 patients, tumor cell DNA shed into their blood showed they had developed relatively high levels of frameshift mutations—and their tumors stopped growing. "It's early days," Diaz cautions, but the results give "a flavor of what we're expecting."

Still, there's an obvious safety concern: The chemotherapy drugs might also create mutations in a patient's healthy cells. Diaz says his group has not seen new tumors in mice treated with the drugs.

Some researchers worry the approach will be counterproductive. They say tumors made up of one or just a few genetically identical cell lineages, or clones, respond better to checkpoint inhibitor drugs than highly heterogenous masses do. Ribas fears that inducing more mutations creates new clones in a tumor and dilutes the impact of any T cells unleashed. He points to a 2019 study in which an Israeli group used UV light to create mutations in melanoma tumors in mice and found the increased diversity of cancer cells actually hampered the checkpoint inhibitor response.

Bardelli says UV light induces a less immunogenic type of mutation than temozolomide. And Diaz and Rousseau argue that their team's two-drug approach will spin out so many potent neoantigen targets for T cells that any harm from greater genetic heterogeneity within the tumor will be minimal.

"What we are doing is different," says Bardelli, who founded a company, with Diaz as an adviser, to develop cancer drugs that will block mismatch repair enzymes—and, the team hopes, turn cancer cells into sitting ducks. ■

IMAGE: STEVE GSCHMEISSNER/SCIENCE SOURCE

EUROPE

Aid keeps researchers afloat in war-torn Ukraine

Fledgling programs aim to sustain lives and work, and stem brain drain

By **Richard Stone**

In early February, when war clouds were gathering over Ukraine, a more personal darkness descended on Yuliia Kosminska. Her 11-month-old daughter, Alyona, tripped at their home—"just a small fall, but she wasn't feeling well," says Kosminska, a thin-film physicist at Sumy State University in the country's northeast. Tests soon revealed an aggressive form of leukemia, and Alyona was transferred to a pediatric cancer clinic in Kyiv. "It was all very quick," she says, her voice catching. On 24 February—the very day the Russians invaded—she got a call with the news she was dreading: Alyona had succumbed to her illness.

Work was a salve for her grief, but it was unsafe to return to Sumy. Kosminska found a glimmer of hope through an emergency aid program run by the Wolfgang Pauli Institute (WPI), a math and physics research center in Austria. The program gives scientists in Ukraine €2000 and Ph.D. students €1500—no strings attached—to sustain their lives and work. Kosminska, one of the first 26 recipients, used some of the money to buy a laptop to prepare an online course on nanomaterials. "You must understand," she says, "I have to do something."

Labs around the world have offered succor to hundreds of Ukrainian scholars who have fled their war-torn country. Far less aid is available for those who, by choice or necessity, have remained. Many have seen their salary payments reduced or stopped as institutes divert resources to the nation's defense. Some have had to flee their homes. They had no one to turn to—until WPI and a few other outfits started to toss lifelines. The sums are small; WPI's initial budget is just €50,000. But the initiatives "are of crucial importance for saving basic and applied science in Ukraine," says Anatoliy Zagorodny, president of the National Academy of Sciences of Ukraine (NASU). Such efforts could help stanch a brain drain that many fear will leave Ukraine depleted long after the war. "We understand that many of these refugees will not return," says Sergiy Ryabchenko of the Institute of Physics.

Grander rescue packages await Ukraine after the war. The U.S. government has begun to discuss the contours of a Marshall Plan, reminiscent of the U.S. effort to rebuild Europe after World War II. Science will not be left out, says Kenneth Myers, president of CRDF Global, a U.S. nonprofit that runs science assistance programs in former Soviet Union nations. "The idea is to not just bring them back to preinvasion levels, but to invest in areas where they can leapfrog and become leaders in science," Myers says.

Kharkiv, Ukraine, 1 week before Russia invaded. Institutes in the city have suffered damage.

PHOTO: AP PHOTO/MSTYSLAV CHERNOV

For now, the focus is survival. Scientists marooned in conflict areas are most at risk, Zagorodny says. With Russian troops mounting a renewed assault on eastern Ukraine, NASU intends to steer €20,000 in aid from the French Academy of Sciences to researchers in eastern science strongholds like Kharkiv and Sumy.

WPI's project is already keeping 17 scientists in those cities afloat. Ihor Shpetnyy, an experimental physicist at Sumy State, stayed behind when fellow residents evacuated. "My task was to find and buy medicine for my parents," he says. His mother's treatment for pulmonary hypertension costs €500 per month. With his salary halved to €220 per month, he says, the WPI money was a lifesaver.

A WPI grant was a career saver for Mykhaylo Mykhaylov. At the end of last year, the solid state physicist had just finished renovating his lab at the Verkin Institute for Low Temperature Physics and Engineering in Kharkiv. Barely 2 months later, with Kharkiv under siege, Mykhaylov escorted his wife and 10-year-old son to the border with Slovakia. He had to say goodbye, because most men under age 60 are barred from leaving Ukraine in anticipation of being called to fight. "We visited a notary, and I made a will. Then I gave my wife all our savings," he says.

Mykhaylov returned to a home wrecked by shelling with €300 in his pocket. "I was considering quitting research," he says. A WPI grant helped him rent an apartment outside Kharkiv—and stick with science. He's writing up a research paper and "eager to return to experimental work, once normality is restored."

Other organizations are joining the rescue effort. The Foundation for Polish Science will fund at least six social science collaborations between scientists in Poland and Ukraine this year; each team will receive €58,000 over 1 year for salaries and research expenses. And the Krzysztof Skubiszewski Foundation, also in Poland, plans to disburse at least €240,000 to scholars in Ukraine and refugees in Poland, with most of it going to those in Ukraine. The French Academy of Sciences is seeking additional funds for NASU from French foundations, and Austria's science ministry just contributed €50,000 to WPI, which will enable it to support another two dozen or so researchers.

The sooner the war ends, the sooner Ukraine can shift from science survival to revival. But first it must endure whatever Russia has in store in the east. In Kharkiv and Sumy, frequent air-raid sirens shatter daily rhythms. "The Russians can come back at any time," says Kosminska, who returned to Sumy with her husband in early April to bury Alyona. Teaching and planning for future experiments help her cope, as does commiserating with friends and colleagues who have also borne hardship and loss. "I can say, for now, I'm OK," she says. ■

ASTROPHYSICS

Dutch dangle bid for gravitational wave detector

Government pledges to support nearly half the cost of the Einstein Telescope

By **Edwin Cartlidge**

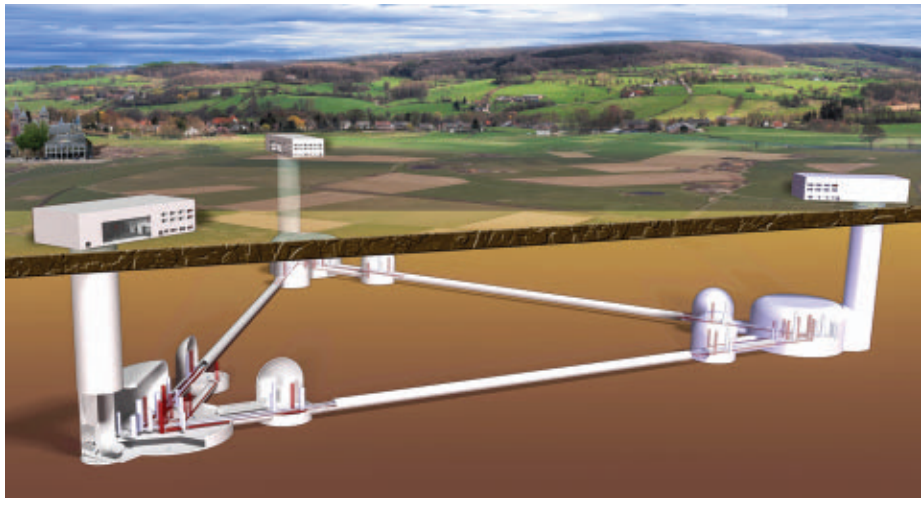
An underground observatory that could detect gravitational waves from the far reaches of the universe is one step closer to reality. Last week, the Dutch government said it was prepared to foot about €900 million of the project's roughly €1.9 billion construction cost if it is built near the border of the Netherlands, Germany, and Belgium. The pledge puts the Dutch proposal for Europe's so-called Einstein Telescope ahead of rival bids, says Stan Bentvelsen, director of the Netherlands's National Institute for Subatomic Physics and a leader of the Dutch proposal. "I think the Dutch government is sticking its neck out the farthest," he says.

Gravitational waves were discovered in 2015 when the ripples in spacetime generated by a pair of merging black holes were detected by the Laser Interferometer Gravitational-Wave Observatory (LIGO), two U.S. detectors each made up of 4-kilometer-long tubes arranged in an L-shape. By sending a laser beam down each tube and bouncing it off mirrors suspended at each end, physicists can look for fluctuations in the beams' travel time, a sign that a passing gravitational wave has subtly stretched or squeezed the arms. Researchers at LIGO and Virgo, a detector in Italy with arms 3 kilometers long, have since gone on to detect dozens more black hole mergers, as well as collisions between pairs of neutron stars.

Now, researchers in Europe and the United States are thinking about bigger detectors that could survey most of the universe (*Science*, 12 March 2021, p. 1089). The most sophisticated, the Einstein Telescope, might be able to detect hundreds of thousands of mergers per year, out to the distant reaches of the observable universe, or soon after the big bang. Three overlapping L-shaped detectors, each with arms 10 kilometers long, will watch for spacetime distortions, and the observatory will be buried in bedrock several hundred meters down to insulate it from the surface noise of wind and traffic. Each detector arm will contain two laser systems, including one cooled almost to absolute zero, giving it sensitivity to longer wavelength radiation from mergers of very large black holes, hundreds of times the mass of the Sun.

Last year, the European Union added the Einstein Telescope to an official wish list of major scientific facilities known as the European Strategy Forum on Research Infrastructures, and official project bids are due within the next 2 to 3 years. Bentvelsen says the Dutch site—somewhere between the cities of Maastricht, Liège, and Aachen—has a layer of soft soil above the bedrock. The contrast between the layers ensures that most surface vibrations would bounce off the bedrock rather than shake the observatory.

In a recent assessment of new proposals for public infrastructure, the Netherlands National Growth Funds Commission cited the region's geology as one of its advantages. Last week, the Dutch Cabinet agreed to pay €42 million for preparatory research and €870 million for construction, should the site be chosen. To win the funding, however, the observatory's backers in academia and provincial government also need to show how the project would benefit local companies. Bentvelsen appears confident, saying he has "positive feedback" from a range of companies and technology institutes in the region.



The Einstein Telescope would be buried several hundred meters underground to insulate it from noise.

ILLUSTRATION: ET CONCEPTUAL DESIGN TEAM

The Dutch bid has an Italian rival that would put the Einstein Telescope close to a former zinc mine near the town of Lula in Sardinia. It is one of the 30 most seismically quiet sites in the world, says Michele Punturo of Italy's National Institute of Nuclear Physics, who co-leads the overall Einstein Telescope collaboration and is the scientific coordinator of the Sardinian proposal. He says he and his colleagues submitted a request in February to the Italian government for €100 million to develop detector technology and better scrutinize the site's geology. They expect to hear the outcome in June. "The noises are positive," he says.

Punturo is heartened that two separate governments are showing an interest in the project, but acknowledges it is not yet guaranteed. After a board of scientists compares site characteristics, the winner should emerge from negotiations between national governments, he says. Construction should begin in 2026 or 2027, with scientific operations starting up some 9 years later.

Even after the big pledge, the Dutch will still have to drum up support from other European nations. Belgium and Germany have been involved in preliminary site investigations but have yet to formally support the Dutch bid.

And Germany may launch its own bid. Günther Hasinger, head of the European Space Astronomy Centre near Madrid, is spearheading plans for a new astrophysics center in the state of Saxony, using federal funds intended for development in eastern Germany, where the coal industry has declined. Hasinger wants to site the Einstein Telescope in a strip of granite found in the region—but only if the site's geology makes it "significantly better than the others." With a site decision expected in 2025, he and his colleagues have a few years to find out. ■

Edwin Cartlidge is a journalist based in Rome.

eppendorf & Science

PRIZE FOR NEURO BIOLOGY

2021 Winner
Amber L. Alhadeff, Ph.D.
Monell Chemical Senses Center, USA
For research on the gut-brain control of hunger circuits

Tell the World About Your Work!

Application Deadline
June 15, 2022

Eppendorf & Science Prize for Neurobiology
The annual Eppendorf & Science Prize for Neurobiology is an international prize which honors young scientists for their outstanding contributions to neurobiological research based on methods of molecular and cell biology. The winner and finalists are selected by a committee of independent scientists, chaired by *Science's* Senior Editor, Dr. Peter Stern. If you are 35 years of age or younger and doing great research, now is the time to apply for this prize.

As the Grand Prize Winner, you could be next to receive
> Prize money of US$25,000
> Publication of your work in *Science*
> Full support to attend the Prize Ceremony held in conjunction with the Annual Meeting of the Society for Neuroscience in the USA
> 10-year AAAS membership and online subscription to *Science*
> Complimentary products worth US$1,000 from Eppendorf
> An invitation to visit Eppendorf in Hamburg, Germany

It's easy to apply! Write a 1,000-word essay and tell the world about your work. Learn more at:

eppendorf.com/prize

eppendorf Science AAAS

AAAS® and Science® are registered trademarks of the American Association for the Advancement of Science, USA. Eppendorf® and the Eppendorf Brand Design are registered trademarks of Eppendorf SE, Germany. All rights reserved, including graphics and images. Copyright © 2022 by Eppendorf SE. Photography: Happy Hour Headshot Philadelphia.

BRIGHT LIGHTS, BIG PITY

City lights lure migratory birds, with lethal results, but weather radar and modeling can help reduce the toll

By **Joshua Sokol**

Every 11 September at dusk, in memory of the 2001 attacks, New York City mounts the *Tribute in Light*, an art installation in lower Manhattan. And every year, as twin towers of light bloom skyward, they attract thousands of migrating birds, sucking in warblers, seabirds, and thrushes—along with predators such as peregrine falcons eager to take advantage of the confusion. On each anniversary, bird conservationists wait below, counting and listening to disoriented chirps. If the observers report too many birds circling aimlessly in the beams, organizers flip off the lights.

In recent years, on-site observers have also used a complementary tool to quantify the orbiting birds: weather radar, which bounces off birds as well as raindrops. In 2017, a group led by Cornell University ornithologist Andrew Farnsworth found that during seven previous anniversaries, the once-a-year installation had attracted a total of about 1.1 million birds. Within 20 minutes of lighting up, up to 16,000 birds crammed themselves into a half-kilometer radius. But when the lights flicked off, the dense clouds of birds on the radar screen dissipated just as fast, a finding later confirmed by on-site thermal cameras.

"That was really illuminating—pun not intended," says that study's lead author, Cornell ecologist Benjamin van Doren. "It really gives you a sense of the scale that light can impact bird migration." The circling birds burn through time and precious body fat, are easy prey, and worst of all, can brain themselves on the windows of nearby buildings.

The finding came as scientists grew increasingly worried about slumping bird numbers. The skies above North America host some 3 billion fewer birds today than in 1970, according to one 2019 analysis. The flocks have faced death by a thousand cuts, including not just light pollution, but climate change, vanishing habitat, and pesticides. Ornithologists fear each added insult could be enough to bend once-abundant bird populations toward extinction (*Science*, 20 September 2019, p. 1228).

The radar studies at the *Tribute in Light* helped lay the groundwork for a tool that could ease the toll: a program Farnsworth's team calls BirdCast, which incorporates continent-scale weather radar and machine learning to forecast the exact nights when hundreds of millions of migratory birds will torrent over U.S. cities. The team then feeds those findings to conservationists and policymakers desperate to help the birds survive the journey by dimming lights along the way.

Such research, paired with existing conservation efforts, is beginning to make a difference: New York City recently passed an ordinance requiring city buildings to turn down lights during migration season, the world's largest city to take such a step. It joins dozens of growing campaigns across the United States that aim to save tens or even hundreds of thousands of birds per city each year. "Given declines over the last 50 years, and what's happening on the planet, everybody needs a win," Farnsworth says.

These warblers and other birds died at the World Trade Center in New York City on 20 September 2021.

LONG BEFORE THOMAS EDISON turned on a light bulb, people were recording anecdotes of nighttime lights ensnaring birds, especially during migration. In 1880, for example, the *Bulletin of the Nuttall Ornithological Club* queried lighthouse keepers across the United States; one respondent, in Florida, reported shielding his face with his hat against an onslaught of "more than a million" birds heading south. He carried away two bushels of crumpled bodies the next morning.

Today's researchers keep more detailed records. In New York City, for example, volunteer crews patrol city sidewalks morning after morning during migration seasons, cataloging birds that thwacked into buildings. Among them are songbirds, light as quarters, that are adapted to every part of their yearly trek between the Amazon and Canada's boreal forests—except for human cities.

"It's really bad to see injured birds that you know are about to die," says Kaitlyn Parkins, a science consultant for New York City Audubon, which has tracked fallen birds since 1997. "I just—I can't internalize it anymore."

Some 3.5 billion birds fly through the southern border area of the contiguous United States each spring, then head north through the center of the country, Cornell ecologist Adriaan Dokter found in a 2018 study using weather radar. Come fall, after breeding season, roughly 4.7 billion birds travel south, this time along routes that cluster more toward the eastern United States. Collisions with buildings take out anywhere from 365 million to 1 billion of them each year, according to one landmark meta-analysis of papers and bird collision data sets published in 2014. That study put buildings behind only domestic cats on the list of top anthropogenic bird killers.

As biologist Pete Marra at Georgetown University puts it, big glass buildings might be more "efficient from an energy perspective [but] are extremely efficient from a killing perspective."

In the daytime, reflective glass can trick birds into flying into what seems like clear, open space, which often ends in death by brain hemorrhage. Overnight, the chief killer is artificial light, which harms birds in several steps. First, when the cumulative glow of light above cities is bright enough to outshine the Milky Way, it sucks in passing migrants from up to 200 kilometers away, a 2018 study using weather radar found. Once in a city, birds are subject to degraded habitats and are prey for housecats, and individual light sources can trap them in pointless, calorie-wasting circles or tempt them headlong into windows. Exposure to light at night has also been linked to disruption of birds' immune systems, microbiomes, foraging behavior, and of course sleep cycles.

To this day, researchers don't know why so many birds find artificial light so alluring. Insect researchers face the same puzzle, debating why, exactly, moths and other insects are drawn to flames and streetlights (*Science*, 7 May 2021, p. 556). "If I could answer that question, I'd be happy," says Kyle Horton, a BirdCast team member now at Colorado State University, Fort Collins. He speculates only that the impulse to fly to light must have once conferred an evolutionary advantage.

WHATEVER THE REASON light lures birds, Farnsworth has toiled since the 1990s to keep them safe from it by tracking and predicting their movements. He aimed to map migration not just as numbers of birds crossing a particular field site, but as flows of biomass operating on regional, even continental

The *Tribute in Lights* installation in New York City lures thousands of migratory birds (bright specks) into twin towers of light.

PHOTOS: MELISSA BREYER; (OPPOSITE PAGE) LAWRENCE LUCIER/FILMMAGIC VIA GETTY

In fall 2020, the city of Dallas turned out lights (left) to spare migratory birds, somewhat dimming the city's typically bright nightscape (right).

scales. On weather radar, multispecies rivers of passing birds can be filtered out from raindrops, allowing their large scale movements to be forecast like weather. (Meteorologists face the same problem—telling apart birds and storm systems—in reverse.) But it took the advent of cloud computing and modern machine learning to crunch through enough data to make BirdCast work.

In 2017, the team did a small-scale test of its methods at the *Tribute in Light*, proving that a single radar station—even one on Long Island, relatively far from the art installation—could quantify how birds move on both large and fine-grained scales. A year later, van Doren and Horton let machine learning models loose on 23 years of continental U.S. weather data. They found peak migration nights regularly involve half a billion birds aloft over the sleeping United States. Given a weather forecast of upcoming temperatures and winds, their model could predict variations in when and where the birds would fly with 62% accuracy 1 week ahead of time, rising to 75% 1 day in advance.

Since then, checking tomorrow's BirdCast has become "common vernacular" for everyone from hobbyists planning birding trips to scientists, says Emily Cohen, a migration ecologist at the University of Maryland Center for Environmental Science. "This real-time forecasting of animal movements is really unprecedented and super-exciting." The Cornell team is working to leverage those predictions to help with the siting of future wind turbines, to forecast spikes in disease transmission from wild birds to farmed poultry, and to anticipate elevated risks of birds smashing into planes, which has bedeviled aviators since the Wright brothers.

At the same time, other teams are using GPS trackers mounted on individual animals to give a bird's-eye view of the same big patterns. One study published this month in the *Journal of Applied Ecology* mapped migration paths in North Africa and Europe, identifying areas where new turbines or power lines would cause collisions.

The birdcasts also reveal where and when migratory flows collide with concentrations of light. In 2019, a study led by Horton ranked the U.S. cities where birds face the greatest collision risk from lights. Chicago, Houston, and Dallas came in on top, as the most light-polluted areas along heavy migration routes (see graphic, p. 344). That information soon percolated to local conservation groups.

Bird advocates have long organized voluntary "Lights Out" campaigns during migration season in a few cities, starting as far back as the 1990s in Toronto, and spreading to at least 44 U.S. cities today. BirdCast data allow conservationists to amplify this message on the exact nights millions of birds are likely to cross an area, using tweets, newsletters, and targeted emails to everyone from bird rehabbers to downtown building managers. "It's wonderful policy influenced by data," Marra says.

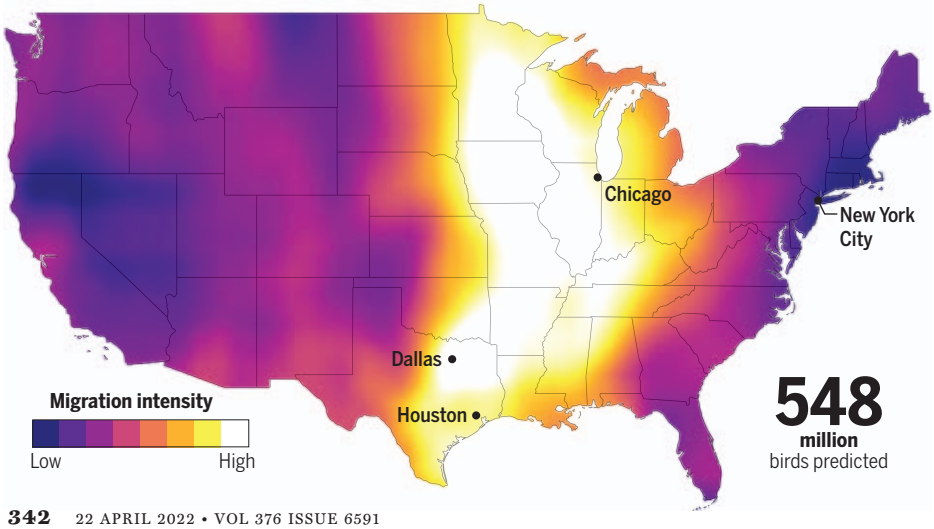
Seeing the ignominious placement of Texan cities on Horton's list—and a Lights Out program already in place in Chicago since the late '90s—the BirdCast group decided to focus its own resources in Texas, joining a smaller campaign that had begun after 400 birds dashed themselves against a single Galveston skyscraper one night in 2017. In 2020, 14 major buildings participated across the state; in 2021, it was more than 100. Most of the iconic skyline buildings in Dallas and Houston turned off nonessential lights last spring and fall for several peak migration weeks during which at least 50% of migrating birds were expected to pass over. Other buildings, for example in downtown Fort Worth, kept their lights off for months to cover the full migratory season. Texas homeowners can sign up for BirdCast alerts that provide narrower night-by-night warnings.

The ability to quantify the risks with radar data has also encouraged advocates to push

Night moves

On peak nights during migration seasons, some half a billion birds fly over the sleeping United States. Based on weather and flight paths, models can forecast where their densities, and peril from city lights, will be highest.

8 September 2021



CREDITS: (PHOTO) LLOYD CLAYTON/CORNELL LAB OF ORNITHOLOGY; (GRAPHIC) C. BICKEL/*SCIENCE*; (DATA) B.M. VAN DOREN AND K.G. HORTON, YEARS OF MIGRATION FORECAST MAP; BIRDCAST MIGRATION FORECAST MAP, 8 SEPTEMBER 2021 20:00 EST. CORNELL LAB OF ORNITHOLOGY; COLORADO STATE UNIVERSITY; UNIVERSITY OF OXFORD

for deeper policy commitments. Newsworthy mass mortality events add to the pressure, such as on 14 September 2021, when heavy migration combined with stormy weather and low cloud cover in New York City led to what Parkins calls a daylong "collision massacre." She says that "to get legislation passed, you have to have the science."

In December 2021, after hearing testimony from Farnsworth and others, the New York City Council voted to require city-owned buildings to turn off lights on peak migration nights, joining several other U.S. cities like Austin, Texas, which passed a similar bill earlier last fall. New York state legislators have also introduced a more aggressive bill that would compel private structures, including residences statewide, to dim or turn off most nonessential lights by 11 p.m. all year. In Illinois, a law passed in summer 2021 requires new state-owned buildings to use bird-safe glass and dimmer, downcast lights shown to reduce collisions; another proposed state bill would ban nonessential lighting during migration season for new buildings near ecological preserves.

THESE MEASURES are incremental, and researchers are eager to learn whether they will help even in the middle of light-polluted urban environments. In one study, published in summer 2021 in the *Proceedings of the National Academy of Sciences*, van Doren and colleagues drew on a unique data set: bird collision records compiled by volunteers from Chicago's Field Museum during 40 years of morning walks around the perimeter of the city's McCormick convention center, on the Lake Michigan waterfront.

From 2000 to 2020, almost 12,000 sparrows, warblers, thrushes, and other birds died at the convention center. The worst single nights left more than 100 bodies behind. Wind conditions and the intensity of migrating birds spotted in archival weather radar seemed to predict how many birds hit the convention center each night, the team found.

But another key factor was how many of the convention center's windows had been illuminated. Each individual bright window left more dead birds for volunteers to find the next day. The correlation suggests halving the number of lit window bays would halve the number of bird strikes, the team estimated, saving thousands of birds at this one three-story building. "It really does seem that each window makes a difference," van Doren says.

Despite the study and the negative press that resulted, the convention center hasn't updated its lighting policy during migration season, says Annette Prince, part of a volunteer group that tracks bird collisions in downtown Chicago. "It's disappointing, because they do have drapes."

Cynthia McCafferty, a spokesperson for the convention center, says the customers who use the building for events dictate the hours the lights stay on. "We have to serve our clients." She adds that closing drapes in the evening would require workers to bring in heavy equipment.

Harder data on the efficacy of Lights Out campaigns are still forthcoming. The BirdCast group, for example, is working with volunteers in Dallas and other Texan sites to build a bigger data set of collisions, as tracked by body counts in the morning. Between 100,000 and 200,000 birds perish in Dallas alone each year, Farnsworth says. Assuming van Doren's findings scale up, their hope is that turning down the city's glow by half could halve that casualty count.

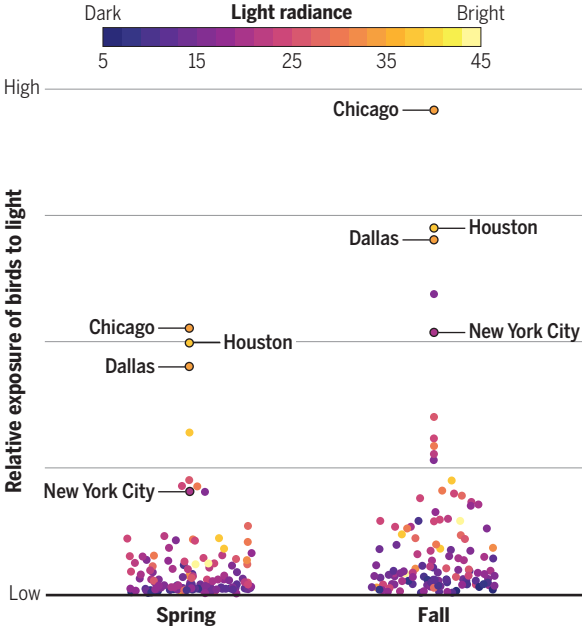
Horton is developing sharper-than-ever migratory forecasts by incorporating information beyond atmospheric conditions, including an area's vegetation and levels of light pollution. He's also identifying public communication strategies to translate them into action. Light pollution harming wildlife is still an unfamiliar concept to many, he says. "Having your porch light on doesn't sound the same as pouring paint down the sewer."

But that means it's also not politically polarized, unlike many environmental topics. "I don't think there's many people out there who are like, 'I want to keep my lights on to kill colorful songbirds,'" Horton says. ■

Joshua Sokol is a science journalist in Raleigh, North Carolina.

Flight risks

By using satellite maps to quantify light at night and radar to estimate the numbers of migratory birds streaming across the night sky, scientists have ranked the cities where birds face the most danger from light pollution. Chicago tops the list in both spring and fall.



CREDITS: (PHOTO) LLOYD CLAYTON/CORNELL LAB OF ORNITHOLOGY; (GRAPHIC) K. FRANKLIN/*SCIENCE*; (DATA) K. HORTON, A. FARNSWORTH *ET AL.*, *FRONTIERS IN ECOLOGY AND THE ENVIRONMENT*, 17, 4, (2019)

INSIGHTS

PERSPECTIVES

PHOTO: MICHEL ROGGO/MINDEN PICTURES

EVOLUTION

A long evolutionary reach for fishing nets

Earlier maturation of Atlantic salmon is linked to indirect effects of fisheries on its prey

By **Nina Overgaard Therkildsen**[1] *and* **Malin L. Pinsky**[2]

Adaptive evolution is not just the stuff of geological history books—it is an ongoing process across ecosystems and can occur on a year-to-year time scale. However, in a world rapidly changing as the result of human activity, it can be challenging to differentiate which changes result from evolution rather than other mechanisms (*1*). On page 420 of this issue, Czorlich *et al.* (*2*) reveal a fascinating example that suggests that commercial fishing drove rapid evolutionary change in an Atlantic salmon population over the past 40 years. Their findings are surprising in two ways—that fishing for salmon drove evolution in the opposite direction from what one would typically expect, and that salmon evolution also was affected by fishing for other species in the ecosystem.

The potential for fishing to drive abnormally fast evolution has been recognized for at least 80 years (*3*), yet conclusive evidence for these effects has remained elusive. The underlying logic is that fisheries often substantially increase adult mortality for harvested populations, creating a fitness benefit for reproduction at earlier ages and smaller sizes. If variation in maturation has a genetic basis, the added mortality from fishing can drive evolution. Laboratory experiments have demonstrated rapid fisheries-induced evolution causing widespread impacts across the fish genome (*4*, *5*), and many wild populations of fish today indeed mature at earlier ages and at smaller sizes compared to decades ago (*6*). Putting lab and field data together, however, has been a major challenge. Atlantic cod, for example, has been a classic putative example of fisheries-induced evolution, yet some of the most heavily fished populations have not revealed clear genomic signals of adaptive evolution (*7*).

Czorlich *et al.* provide some of the strongest evidence to date linking fishing to changes in the genetic composition and maturation of a wild population. They studied Atlantic salmon, which are born in fresh water, migrate to the ocean to grow,

Czorlich *et al.* have uncovered a link between the intensity of fishing activities and a gene associated with maturation age in Atlantic salmon.

and then return to their birthplace once they are mature and ready to spawn. There is substantial variation in the time salmon spend at sea, both within and across populations, and many populations now mature and return to spawn at an earlier age (*8*).

In previous studies, researchers had identified a single gene—the *vestigial-like family member 3* (*vgll3*) gene—that has a large effect on salmon maturation timing (*9*). By analyzing the DNA collected from scales of 1319 salmon captured between 1975 and 2014 from the Tenojoki population in the Teno River located at the Finland-Norway border (*10*), researchers also uncovered that a shift toward earlier maturation over this period coincided with an 18% increase in the frequency of a *vgll3* variant linked with early maturation (*10*). To identify the drivers behind this change, Czorlich *et al.* compared interannual fluctuations in the frequency of early- and late-maturing genetic variants with historical data on environmental, ecological, and fishing-related factors. They revealed that temporal genetic changes were associated with fishing, consistent with fisheries-induced evolution, but in surprising ways. Rather than selecting for earlier maturation, as typically expected, salmon fishing in the Teno River had selected for later maturation. This result highlights the subtle context dependence of evolution. In this case, it appears to have emerged because the main fishing method in this river primarily captures smaller, younger fish and leaves older, larger fish to spawn.

Perhaps even more surprising, another major factor that strongly correlated with the genetic changes in salmon was the population size of a small marine fish called capelin. During their ocean development, salmon grow in part by feeding upon capelin. But capelin have experienced marked population booms and busts over the last few decades, in large part because of capelin fishing. Capelin appear to influence salmon evolution through the at-sea survival of salmon, which is lower when fewer capelin are present. With lower survival, early- rather than late-maturing salmon are especially likely to return to their natal population and pass their genes on to the next generation, increasing the number of early-maturing salmon. Thus, capelin fisheries affected not just their target species, but had cascading evolutionary effects elsewhere in the ecosystem. In an interesting coincidence, the capelin are caught partly as feed for farmed Atlantic salmon, revealing distant and unexpected impacts of fish farming. These ripple effects from fisheries and aquaculture have been widely appreciated for ecological processes, but Czorlich *et al.* show a rare example of evolutionary impacts beyond the targeted species.

The distinct large-effect gene region that accounts for 40% of the variation in Atlantic salmon maturation timing made it feasible for Czorlich *et al.* to track adaptive evolution at the genetic level. Large-effect loci have also been discovered in other fish species—e.g., affecting the migration timing in Pacific salmonids (*11*) and the growth rate in Atlantic silversides (*5*)—but not all traits are like this. Many growth and maturation traits are expected to be highly polygenic—that is, influenced by small effects from hundreds of different genes (*12*). Changes in traits with such genomic architectures will be more difficult to detect.

Another key to Czorlich *et al.*'s discovery was the detailed series of annual allele frequency estimates from the Tenojoki population that allowed the testing of environmental associations not just based on long-term change, but with year-to-year fluctuations. The inference from Czorlich *et al.* remains correlational and does not provide definitive evidence of causation, but the high temporal resolution provides great power to test associations, highlighting the value of comprehensive historical collections.

Despite the clarity of the Tenojoki salmon example, many questions remain. For instance, the evolutionary result does not appear immediately generalizable to other salmon populations. Despite exposure to some of the same fisheries and capelin interactions, a different Teno River salmon population does not show the same evolutionary change toward earlier maturation, perhaps because the effects of *vgll3* on maturation are mediated by other, as-yet-unknown genes (*10*). The broader genomic footprint of fishing also remains an important and open question, because the study only examined a couple hundred of the roughly 3 billion positions in the salmon genome.

Whether and how human activities drive rapid evolutionary change, such as how they influence ecosystems beyond the directly affected species, remains a vital research topic. Evolutionary impacts may be widespread but poorly recognized because they have been hidden from view. A key question will be whether evolutionary impacts can be predicted, perhaps with better knowledge of genomic architecture. Answering these questions is not just an academic exercise. The effects of fisheries-induced evolution can include lower population productivity and greater population instability, and can be difficult to reverse (*6*). The first step toward mitigating their negative impacts will be understanding when and where they occur. ■

[1]Department of Natural Resources and the Environment, Cornell University, Ithaca, NY, USA. [2]Department of Ecology, Evolution, and Natural Resources, Rutgers University, New Brunswick, NJ, USA. nt246@cornell.edu

REFERENCES AND NOTES

1. F. Pelletier, D. W. Coltman, *BMC Biol.* **16**, 7 (2018).
2. Y. Czorlich *et al.*, *Science* **376**, 420 (2022).
3. J. B. S. Haldane, *Am. Nat.* **71**, 337 (1937).
4. S. Uusi-Heikkilä, T. Sävilammi, E. Leder, R. Arlinghaus, C. R. Primmer, *Mol. Ecol.* **26**, 3954 (2017).
5. N. O. Therkildsen *et al.*, *Science* **365**, 487 (2019).
6. M. Heino, B. Diaz Pauli, U. Dieckmann, *Annu. Rev. Ecol. Evol. Syst.* **46**, 461 (2015).
7. M. L. Pinsky *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2025453118 (2021).
8. J. Erkinaro *et al.*, *Can. J. Fish. Aquat. Sci.* **76**, 42 (2019).
9. N. J. Barson *et al.*, *Nature* **528**, 405 (2015).
10. Y. Czorlich, T. Aykanat, J. Erkinaro, P. Orell, C. R. Primmer, *Nat. Ecol. Evol.* **2**, 1800 (2018).
11. N. F. Thompson *et al.*, *Science* **370**, 609 (2020).
12. N. H. Barton, P. D. Keightley, *Nat. Rev. Genet.* **3**, 11 (2002).

ACKNOWLEDGMENTS

N.O.T. acknowledges funding from NSF OCE-1756316. M.L.P. acknowledges funding from NSF OISE-1743711 and serves as an Oceana science adviser.

10.1126/science.abo6512

CELL BIOLOGY

Resisting attack by repairing the damage

Removing membrane pores may help cancer cells survive T cell assault

By **Norma W. Andrews**

Tumor cells or cells infected by pathogens are thought to resist killing by cytotoxic T lymphocytes (CTLs), but the mechanisms involved are incompletely understood. On page 377 of this issue, Ritter *et al.* (*1*) report that the membrane-remodeling endosomal sorting complexes required for transport (ESCRT) proteins accumulate at cytolytic synapses, the specialized sites where CTLs bind target cells and secrete the pore-forming protein perforin. Using high-resolution imaging and a clever strategy for selecting cells at early stages of permeabilization, they show that ESCRT protein recruitment coincides with the onset of injury and helps cancer cells resist CTL attack in vitro.

ESCRT proteins drive membrane remodeling and scission events in various processes, including viral budding and cytokinesis (*2*, *3*). The earlier discovery that they facilitate wound resealing (*4*) suggested a direct path for lesion removal. ESCRT proteins were detected at plasma membrane sites wounded by lasers (*4*, *5*), but uncertainties remained because lasers can cause heat-related artifacts and denature membrane components. Ritter *et al.* settle this issue by visualizing ESCRT proteins precisely at the site of a highly physiological form of membrane injury. They also observe membrane protrusions containing ESCRT proteins within the cytolytic synapse, consistent with a previously proposed mechanism of membrane repair by vesicle shedding (*4*, *5*). However, the membrane detachment events observed to date could have resulted from laser-induced damage or cell death. Consistent with this possibility, Ritter *et al.* only directly observed membrane shedding in cells that failed to reseal after CTL attack. Thus, how plasma membrane lesions are ultimately removed remains unclear.

Ca^{2+} influx through cell membrane wounds triggers the formation of micrometer-scale blebs as a consequence of cortical cytoskeleton rearrangements (*4*, *6*). These large blebs, which display a high density of membrane-associated ESCRT proteins (*4*), do not promote lesion removal because the blebs retract back into the cell body when membrane integrity is restored (*4*, *6*), and cells reseal normally when blebbing is prevented (*6*). It is therefore important, in the context of CTL–target cell engagement, to distinguish ESCRT-containing membrane buds that are responsible for shedding perforin lesions from reversible membrane perturbations caused by Ca^{2+} entry. One strategy might be to compare the topography of cytolytic synapses resulting from CTL–target cell engagements of variable potency, as described by Ritter *et al.*, to identify morphological features specifically associated with more-efficient ESCRT-dependent membrane repair. Would target cells engaged by CTLs under conditions that favor ESCRT-dependent survival show more-homogeneous, perhaps smaller ESCRT-containing membrane buds within the cytolytic synapse?

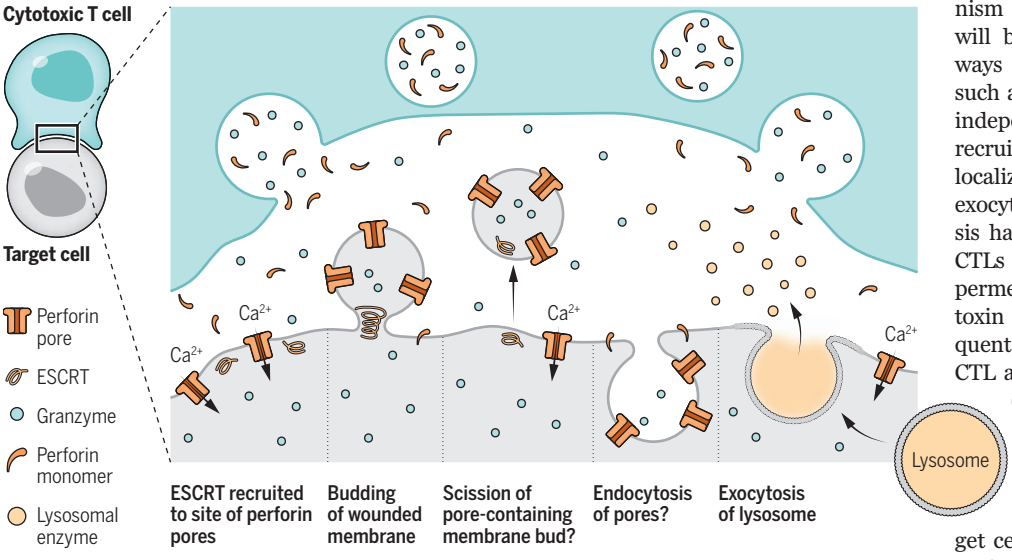
In addition to understanding the mechanism of lesion removal, another challenge will be to determine whether other pathways linked to plasma membrane repair, such as exocytosis and endocytosis (*7*), work independently of or in concert with ESCRT recruitment. Although not with the precise localization shown by Ritter *et al.* for ESCRT, exocytosis of lysosomes and rapid endocytosis have been observed in cells attacked by CTLs (*8*), mimicking observations in cells permeabilized by the bacterial pore-forming toxin streptolysin O (SLO) (*6*, *9*). A subsequent study challenged the conclusion that CTL attack enhances target cell endocytosis (*10*), but this discrepancy may be explained by the different endocytosis markers used in each study.

Granzyme molecules responsible for the CTL lethal hit can enter target cells through perforin pores assembled at the cell surface (*10*). However, it remains possible that after granzyme entry, perforin pores are rapidly internalized. Perhaps this explains why Ritter *et al.* could not detect pore-like structures on the surface of target cells by focused ion beam–scanning electron microscopy (FIB-SEM). It will be interesting to extend the use of FIB-SEM to examine the endosomal compartment of recently permeabilized CTL target cells to see whether it undergoes the expansion observed by others (*6*, *8*, *9*) and whether it contains structures that could be attributed to internalized perforin pores.

In addition to ESCRT proteins, others have detected exocytosis of lysosomes at

Membrane wounding and repair at the cytolytic synapse

Cytotoxic T cells secrete perforin and granzymes, which trigger target cell death. Target cells resist attack through endosomal sorting complexes required for transport (ESCRT) proteins, which may promote shedding of perforin pores. This process may also involve Ca^{2+}-triggered exocytosis of lysosomes and rapid endocytosis.



GRAPHIC: A. MASTIN/*SCIENCE*

Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD, USA.
Email: andrewsn@umd.edu

wound sites and suggested that an exocytosis-mediated local drop in membrane tension might recruit ESCRT proteins (*4*). Thus, plasma membrane repair may involve the concerted action of Ca^{2+}-regulated lysosomal exocytosis, up-regulation of endocytosis, and ESCRT recruitment (see the figure). This scenario is consistent with the frequent detection of rapid Ca^{2+}-dependent lysosomal exocytosis in cells permeabilized by various mechanisms (*7*, *11*, *12*) and with the finding that the inhibition of lysosomal exocytosis blocks plasma membrane repair (*7*, *13*). Notably, the lysosomal enzyme acid sphingomyelinase, when released extracellularly by wounded cells, can be sufficient for promoting endocytosis and plasma membrane repair (*9*). With the exciting possibilities now offered by advanced imaging techniques, the next step should be to define the spatiotemporal relationship of membrane wounding with lysosomal exocytosis, endocytosis, and ESCRT recruitment. If this can be achieved, it would greatly facilitate the challenging process of deciphering how each of these pathways contributes to lesion removal. This is important because a comprehensive understanding of the plasma membrane resealing mechanism may reveal steps amenable to therapeutic intervention, as illustrated by the recent report of enhanced muscle repair by in vivo expression of lysosomal acid sphingomyelinase (*14*).

ESCRT proteins also promote repair of nanoscale lesions in the endolysosomal network and the nuclear envelope, by mechanisms that are still poorly understood but that may not involve vesicle budding (*15*). Surprises certainly lie ahead as the fascinating roles of ESCRT proteins in membrane remodeling are better understood along with their relationship, if any, to wound-induced membrane trafficking events. ■

REFERENCES AND NOTES

1. A. T. Ritter *et al.*, *Science* **376**, 377 (2022).
2. M. Vietri, M. Radulovic, H. Stenmark, *Nat. Rev. Mol. Cell Biol.* **21**, 25 (2020).
3. J. Schöneberg, I.-H. Lee, J. H. Iwasa, J. H. Hurley, *Nat. Rev. Mol. Cell Biol.* **18**, 5 (2017).
4. A. J. Jimenez *et al.*, *Science* **343**, 1247136 (2014).
5. L. L. Scheffer *et al.*, *Nat. Commun.* **5**, 5646 (2014).
6. V. Idone *et al.*, *J. Cell Biol.* **180**, 905 (2008).
7. N. W. Andrews, M. Corrotte, *Curr. Biol.* **28**, R392 (2018).
8. D. Keefe *et al.*, *Immunity* **23**, 249 (2005).
9. C. Tam *et al.*, *J. Cell Biol.* **189**, 1027 (2010).
10. J. A. Lopez *et al.*, *Blood* **121**, 2659 (2013).
11. C. S. Ferranti *et al.*, *J. Cell Biol.* **219**, e201903176 (2020).
12. C.-L. Forestier, C. Machu, C. Loussert, P. Pescher, G. F. Späth, *Cell Host Microbe* **9**, 319 (2011).
13. M. Encarnação *et al.*, *J. Cell Biol.* **213**, 631 (2016).
14. D. C. Bittel *et al.*, *J. Clin. Invest.* **132**, e141295 (2022).
15. K. P. Bohannon, P. I. Hanson, *Curr. Opin. Cell Biol.* **65**, 122 (2020).

ACKNOWLEDGMENTS

N.W.A. has consulted for Casma Therapeutics, Inc.

10.1126/science.abp8641

PHOTO: HDOT

A dynamic message sign displays traffic deaths above the H-1 Freeway on the Liliha Street overpass in Hawaii.

BEHAVIORAL SCIENCE

How safe are safety messages?

Highway fatalities increased in response to certain messages

By **Gerald Ullman** *and* **Susan Chrysler**

With 1.35 million people killed in road crashes worldwide each year (*1*), highway agencies search for ways to reduce traffic deaths and injuries, including encouraging safer driving behaviors. Electronic dynamic message signs (DMSs) (see the photo) are viewed as highly effective devices for communicating traffic safety messages directly to the driving public with the goal of improving road safety (*2*). However, the actual effects of such DMS messages on road safety have never been evaluated in a rigorous manner. On page 370 of this issue (*3*), research by Hall and Madsen suggests that, contrary to expectations, displaying traffic fatality numbers in traffic safety messages on DMSs is associated with an increase in crashes downstream.

Beginning in 2012, the Texas Department of Transportation (TxDOT) began posting traffic safety messages for 1 week per month on its statewide network of nearly 900 DMSs. Messages consisted of a traffic safety slogan (e.g., DON'T DRINK AND DRIVE) and the cumulative number of traffic fatalities that year on Texas roadways (XXX TRAFFIC DEATHS ON TEXAS ROADWAYS IN 2012). These messages were displayed when the signs were not used for conveying other transportation-related information about incidents, roadwork, or special events. The rest of the month, fatality information was not displayed.

Texas A&M Transportation Institute, College Station, TX, USA. Email: g-ullman@tti.tamu.edu

Hall and Madsen compared crashes downstream of DMSs across the state during periods when traffic safety messages with fatality numbers were being displayed versus not being displayed. To control for other possible factors influencing their results, they also compared crashes on those same roadway segments before the fatality message campaign began and on roadway segments upstream of the DMSs. They concluded that the display of traffic safety messages with fatality numbers resulted in a 1.35% increase in traffic crashes up to 10 km downstream of the DMSs.

Hall and Madsen contend that these results suggest that messages with fatality numbers are overly salient to drivers. They do not discuss the valence of emotions that fatality messages induce, but instead focus on their salience in the working memory of drivers leading to cognitive distraction, which leads to driving errors. The finding seems inconsistent with other research that has found that the use of fatalities and other statistics in traffic safety campaigns is mostly ineffective in influencing driver attitudes or behaviors, in large part because of "optimism biases" held by most drivers regarding their abilities to operate a vehicle and avoid becoming involved in a fatal crash (*4*, *5*). However, because the effect of the fatality messages was greater in urban areas, the issue may be one of excessive salience or of some cognitive overload. Given the greater baseline cognitive demand of multilane urban freeways compared with rural highways, the additional cognitive load induced by fatality messages may be enough to push

some drivers beyond their attentional capacity. Such effects of attentional overload have been demonstrated in driving simulations, naturalistic driving studies, and closed-course evaluations (*6*).

Research on the effects of emotional distraction upon the driving task is limited, but there are a few studies that support the hypothesis. This research often relies on self-report and surveys of past driving incidents, making causal attributions difficult. Laboratory and simulation studies must attempt to induce a particular emotion to study its effect. One simulator study of the effects of auditory messages of positive and negative emotion (*7*) showed that negative words reduced driving speed and worsened lateral control compared with positive words. Evoked response potentials used to assess the allocation of attentional resources across tasks showed that positive and negative stimuli were processed differently. Similar studies suggest that negative billboard images can degrade driving performance (*8*, *9*). One simulator study showed that induced happiness and anger each caused more driving errors compared with neutral and fear conditions, but subjective workload was similar across the affective states (*10*). Conversely, other studies suggest that drivers will modulate their glances to billboards based on the situational demands of the driving task (*11*, *12*). Whether the same modulation occurs for messages presented on DMSs during times of high attentional demand is unknown.

Another plausible hypothesis for the results obtained by Hall and Madsen is that it is the overall design of the traffic safety messages, including fatality numbers, which collectively contributes to an information overload situation that has adverse effects upon driving behavior. Messages must be limited in length and formatted to ensure that motorists can quickly read and correctly process the information presented during limited viewing time. Guidelines and regulations have been developed on how to best design DMS messages reporting things like traffic incidents, special events, and roadwork activities (*13*). Similar guidance does not yet exist for traffic safety messages. These messages are often unclear in terms of how drivers should respond to the information. It has commonly been assumed that drivers simply read and then quickly disregard messages that they deem unnecessary. However, the results of Hall and Madsen suggest that drivers may continue to try and assess how they are supposed to use that information for a much longer period of time after reading the message.

Although not something that Hall and Madsen could explicitly test with the Texas dataset, this hypothesis would help explain why a message containing fatality numbers could impede a driver's cognitive abilities and adversely affect their driving performance but not influence their attitudes or conscious driving behaviors. This hypothesis would also suggest that similar effects would be expected when using other numbers in traffic safety messages, such as the number of speeding tickets issued, the percentage of crashes involving impaired motorists, etc. (again, a hypothesis that Hall and Madsen could not test with the existing Texas dataset). Although not necessarily in response to the Hall and Madsen results, it should be noted that the US Federal Highway Administration in 2021 discouraged the use of fatality numbers and other statistics in traffic safety messages displayed on DMSs (*14*).

The crash data presented by Hall and Madsen clearly demonstrate a safety effect of showing fatality numbers on DMSs. However, the mechanism for this safety effect is not clearly elucidated by the data presented in the paper. Additional analyses regarding crash types and documented causal factors in the crash reports might yield more insights. For example, the authors treated all types of crashes as equal and only separated single-vehicle from multiple-vehicle crashes. The assertion that emotional salience caused distraction would predict a pattern of crash types that would be the result of distraction, such as rear-end crashes resulting from delayed response to a slowing lead vehicle. Examining the pattern of specific crash configurations would be a stronger test of the distraction explanation posited by the authors. ■

REFERENCES AND NOTES

1. National Center for Injury Prevention and Control (NCIPC), "WISQARS™ – Web-based Injury Statistics Query and Reporting System" [Centers for Disease Control and Prevention (CDC), 2018]; www.cdc.gov/injury/wisqars.
2. Federal Highway Administration (FHWA), "Manual on Uniform Traffic Control Devices (MUTCD)" [US Department of Transportation (US DOT), 2009].
3. J. D. Hall, J. Madsen, *Science* **376**, eabm342 (2022).
4. B. Elliot, "Road Safety Mass Media Campaigns: A Meta Analysis," report no. CR 118 (Federal Office of Road Safety, Australia, 1993).
5. N. Guttman, *Accid. Anal. Prev.* **84**, 153 (2015).
6. M. A. Regan, J. D. Lee, K. L. Young, Eds., *Driver Distraction: Theory, Effects, and Mitigation* (CRC Press, 2008).
7. M. Jeon, B. N. Walker, J.-B. Yim, *Transp. Res. Part F Traffic Psychol. Behav.* **24**, 197 (2014).
8. H. E. K. Walker, L. M. Trick, *Safety Sci.* **115**, 121 (2019).
9. A. Megías, A. Maldonado, A. Cándido, A. Catena, *Accid. Anal. Prev.* **43**, 813 (2011).
10. M. Chan, A. Singhal, *Safety Sci.* **72**, 302 (2015).
11. K. L. Young, A. N. Stephens, D. B. Logan, M. G. Lenné, *Appl. Ergon.* **60**, 136 (2017).
12. J. S. Decker *et al.*, *Traffic Inj. Prev.* **16**, 234 (2015).
13. C. L. Dudek, "Changeable Message Sign Operation and Messaging Handbook," report no. FHWA-OP-03-070 (FHWA, US DOT, 2004).
14. M. R. Kehrli, "Uses of and Non-Standard Syntax on Changeable Message Signs," MUTCD Official Ruling no. 2(09)-174(I) (FHWA, US DOT, 2021).

10.1126/science.abq1757

ELECTROCHEMISTRY

Electrifying membranes to deliver hydrogen

An electrochemical membrane reactor enables efficient hydrogen generation

By **Arthur J. Shih** *and* **Sossina M. Haile**

The developed world has had a vacillating interest in hydrogen (H_2) as the green fuel of the future. Today, the interest is being renewed as the climate crisis becomes increasingly evident (*1*). A key challenge with hydrogen, presuming that it can be generated by using sustainable electrical power, is its economical delivery. The daunting cost of installing a hydrogen infrastructure has been a major driver behind the decision of policy-makers in the United States and elsewhere to put the hydrogen effort on hold (*2*). On page 390 of this issue, Clark *et al.* (*3*) address head-on the hydrogen infrastructure need by exploiting electrochemical membrane reactors to strip hydrogen from more convenient carriers, including ammonia (NH_3), methane (CH_4), and biomass. These fuels could potentially be delivered to a point of need by using an existing infrastructure, where they could then be converted to hydrogen for use in fuel cells.

The concept of using liquid or easily liquified hydrogen carriers to fulfill hydrogen delivery needs has gained traction in recent years (*4*–*6*). Ammonia as the carrier is attractive because the cycle is entirely carbon free; whereas methane is attractive because the locally produced carbon dioxide can potentially be sequestered; and biomass is attractive because if deployed alongside sequestration, it results in a carbon-negative cycle. Among the reactor types available for extracting hydrogen from hydrogen-bearing compounds, electrochemical membrane reactors based on proton ceramic electrolytes offer distinct advantages. Such reactors combine thermochemical

Materials Science and Engineering, Northwestern University, Evanston, IL 60208, USA. Email: sossina.haile@northwestern.edu

catalysts that facilitate decomposition of the carrier with electrochemical pumping of hydrogen across a proton-conducting solid-state membrane (*5*, *7*). Because only protons, which emerge in the form of hydrogen gas upon undergoing oxidation at the hydrogen evolution electrode, can be delivered across the membrane, one can reasonably anticipate that every electron delivered to the membrane will result in the production of hydrogen in a 1:2 ratio. Furthermore, because of the solid-state and gas-impermeable nature of the membrane, one can expect the hydrogen produced to be entirely free of impurities—in particular, of unreacted carrier molecules, species that simply cannot get to the other side of the membrane. Another added benefit is the ability to pressurize the hydrogen by only increasing the current.

As water electrolysis for hydrogen fuel production ramps up, the energy sector is in increasing need of efficient ways to deliver that hydrogen using convenient carriers such as ammonia and methane.

PHOTO: YOSHIKAZU TSUNO/AFP/GETTY IMAGES

In contrast to their electrochemical counterparts, traditional catalytic membrane reactors use a membrane that is hydrogen-permeable, typically made with palladium (Pd) or a Pd alloy, rather than one that is proton-permeable (*8*–*15*). Hydrogen is driven across a traditional membrane by mechanical pressure, which creates a chemical potential gradient. In the electrochemical membrane reactor, protons are driven across the membrane by application of a voltage (or current), which indirectly drives the flux of hydrogen gas.

Recent advances in electrochemical membrane reactors (*5*, *7*) have spurred the race to implement "hydrogen-on-demand" solutions. In devices scaled up for practical applications, challenges emerge. As noted by Clark *et al.*, managing the temperature profile across the reactor is particularly difficult. The process of pumping hydrogen across an electrochemical membrane leads to an increase in temperature because of the changes in hydrogen concentrations. At the same time, the decomposition reactions are inherently endothermic and drive the temperature down. Consequently, in a reactor with a simple linear flow, the upstream regime will be much cooler than the downstream regime. Such a temperature gradient introduces efficiency penalties.

Clark *et al.* meet this challenge by engineering a counterflow geometry that enables transfer of the heat generated at the downstream portion of the reactor, as a consequence of the electrochemical pumping, to the upstream portion of the reactor, where the carrier decomposition reactions cool the system (*3*). Beyond the use of a counterflow design, thermal gradients are mitigated by formulating an interconnect material that provides excellent heat transfer as well as electrical contact between adjacent cells in the reactor. The interconnect composition is also designed to match the thermal expansion behavior of the electrochemical components of the reactor, contributing to its long-term stability. With these advances in reactor design and material components, the authors achieved an unprecedented combination of carrier gas conversion, hydrogen recovery, system size, and reactor lifetime.

The >99% hydrogen extraction efficacy of the system of Clark *et al.* exceeds all other values in the literature. Although extraction efficacy is not a commonly discussed metric, it is useful for describing the overall performance of a catalytic membrane reactor and can be calculated by multiplying the carrier conversion fraction by the hydrogen recovery fraction. Another important metric is the pressure difference across the membrane. In traditional catalytic membrane reactors, in which mechanical pumps pressurize the reactant supply, the permeate emerges at a pressure lower than that of the feed. Therefore, additional mechanical pumps are required to pressurize and compact the hydrogen for storage and transport. Clark *et al.* demonstrated an integrated system in which chemical transformation, purification, and pressurization are all achieved in a single device, an accomplishment that is only possible in an electrochemical membrane reactor. The combination of hydrogen extraction efficacy and exhaust gas pressurization achieved in their system are truly unprecedented. Future efforts will likely be directed toward increasing the hydrogen flux, which remains moderate for their electrochemical system and does not factor into the extraction efficacy or pressurization metrics.

Today, the main application of hydrogen is in oil refining, which accounts for about 55% of all hydrogen consumption, and about 93% of hydrogen is produced by using methane (*1*). Consequently, technological advances in methane steam reforming may inadvertently prolong the global reliance on fossil fuels. By directing greater attention to ammonia, one of the hydrogen carriers demonstrated by Clark *et al.*, future electrochemical catalytic reactors may allow use of hydrogen without incurring carbon emissions. ■

REFERENCES AND NOTES

1. T. Gul, D. Turk, *The Future of Hydrogen: Seizing today's opportunities* (International Energy Agency, 2019).
2. M. Wald, "U.S. Drops Research Into Fuel Cells for Cars," New York Times, 7 May 2009.
3. D. Clark, *Science* **376**, 390 (2022).
4. W. C. Leighty, J. H. Holbrook, J. P. International, in *19th World Hydrogen Energy Conference* (WHEC). (Elsevier, 2012), vol. 29, pp. 332–345.
5. D. K. Lim *et al.*, *Joule* **4**, 2338 (2020).
6. S. Chatterjee, R. K. Parsapur, K.-W. Huang, *ACS Energy Lett.* **6**, 4390 (2021).
7. H. Malerød-Fjeld *et al.*, *Nat. Energy* **2**, 923 (2017).
8. H. W. Abu El Hawa, S. N. Paglieri, C. C. Morris, A. Harale, J. Douglas Way, *Separ. Purif. Tech.* **147**, 388 (2015).
9. V. Cechetto *et al.*, *Fuel Process. Technol.* **216**, 106772 (2021).
10. J. L. Cerrillo *et al.*, *Chem. Eng. J.* **431**, 134310 (2022).
11. B. Dittmar *et al.*, *Int. J. Hydrogen Energy* **38**, 8759 (2013).
12. C.-H. Kim *et al.*, *Int. J. Hydrogen Energy* **43**, 7684 (2018).
13. T. Kume *et al.*, *J. Nat. Gas Sci. Eng.* **3**, 591 (2011).
14. J. A. Medrano *et al.*, *Int. J. Hydrogen Energy* **41**, 8706 (2016).
15. M. Sarić, Y. C. van Delft, R. Sumbharaju, D. F. Meyer, A. de Groot, *Catal. Today* **193**, 74 (2012).

10.1126/science.abo5369

SUPERCONDUCTIVITY

Superconductors gain momentum

Spin-density modulations point to inhomogeneous superconductivity in a perovskite

By **Eva Pavarini**

In a superconducting material, electrical resistivity abruptly disappears below a critical temperature. Discovered in solid mercury in 1911, superconductivity remained an unsolvable riddle until 1957, when physicists Bardeen, Cooper, and Schrieffer developed a theory explaining the phenomenon (*1*). According to the Bardeen-Cooper-Schrieffer (BCS) scheme, superconductivity arises when electrons form pairs that behave in a way that allows current to flow with zero resistance. Then, in 1964, Fulde and Ferrell (*2*) and Larkin and Ovchinnikov (*3*) pointed out that in the presence of a magnetic field, a different type of superconducting electron pairs could form. However, despite the intense search, direct evidence of this Fulde-Ferrell-Larkin-Ovchinnikov (FFLO) superconducting state has proven hard to find. On page 397 of this issue, Kinjo *et al.* (*4*) report the observation of FFLO-driven spin-density modulations in the layered perovskite Sr_2RuO_4—a system with its own peculiar history.

The findings of Kinjo *et al.* signal the presence of inhomogeneous superconductivity, the hallmark of the FFLO state—a superconducting state that had remained elusive in the long search for its existence and can offer an alternate route to explore the diversity of superconductivity. To understand what makes it peculiar, we have to discuss first the standard case. Superconductors whose property can be explained with the BCS framework are called "conventional superconductors." When superconductivity arises in these materials, the electron pairs are made of opposite-spin electrons through a mechanism known as spin-singlet pairing. Moreover, the electron pairs carry no momentum, making the superconducting state homogeneous. In a conventional superconductor, the presence of an external magnetic field can destabilize the pairs because electrons with spin parallel and antiparallel to the magnetic field acquire different energy. If the resulting energy difference, known as Zeeman splitting, is sufficiently large, the magnetic field can unpair the electrons, and the material will go back to its normal, nonsuperconducting state.

However, under certain conditions, superconductivity could survive this situation. Electrons could form an unusual kind of electron pairs, carrying a nonzero momentum (see the figure). This type of pairing creates the FFLO state, characterized by the spatial modulations arising from the momentum of the pairs. However, the FFLO state is difficult to induce and observe in materials because it can be easily destabilized. Its realization is favored if the superconductor is in the clean limit, a condition in which electrons can move without being scattered for sufficiently long distances. Even more important for the FFLO state to emerge, the Zeeman splitting must be the main mechanism that would otherwise destroy the superconductivity in the material. In most superconductors, however, other pair-breaking mechanisms are stronger, and the FFLO state has no chance of forming.

Two-dimensional systems with heavy charge carriers, thanks to their sensitivity to Zeeman splitting, are good candidates for the search for FFLO states. Signatures of the FFLO state have been reported in layered heavy fermions (*5*, *6*) and certain two-dimensional organic materials (*7*, *8*). With these considerations in mind, the layered perovskite Sr_2RuO_4 investigated by Kinjo *et al.* possesses the basic qualities of an FFLO system: It is in the clean limit, is layered, and (*9*) has charge carriers with large effective mass. For a long time, however, Sr_2RuO_4 was believed to be a chiral spin-triplet superconductor (*10*, *11*) and therefore a type of system that cannot possess an FFLO state. This is because spin-triplet superconductors have electron pairs whose spins point in the same direction, and thus the Zeeman splitting cannot break them.

The prospect of having at hand a true realization of spin-triplet superconductivity—a very rare phenomenon—has put Sr_2RuO_4 at the center of very intense investigations. With time, it became clear that not all experimental observations on Sr_2RuO_4 are consistent with the chiral spin-triplet description. Among the increasing collection of unexplained observations was a discontinuous transition from superconducting to the normal state with an increasing magnetic field (*12*). In a spin-singlet superconductor, this behavior could be explained by the Zeeman splitting breaking the pairs. In the case of Sr_2RuO_4, however, such an explanation was at odds with the accepted chiral spin-triplet picture.

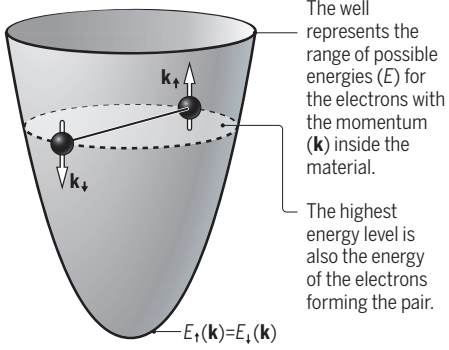
Initially, rather than questioning established view, the puzzling measurement prompted the search for alternative explanations for these inconsistencies while preserving the chiral spin-triplet interpretation. However, the consensus began to shift in the past few years, when more clear

Electron pairs with a nonzero momentum

In conventional superconductivity, electron pairs have zero overall momentum. However, in the presence of a magnetic field, electron pairs can gain a nonzero momentum because of the different energies acquired by opposite-spin electrons.

Bardeen-Cooper-Schrieffer electron pair

Momentum: $\mathbf{k}_\uparrow + \mathbf{k}_\downarrow = 0$

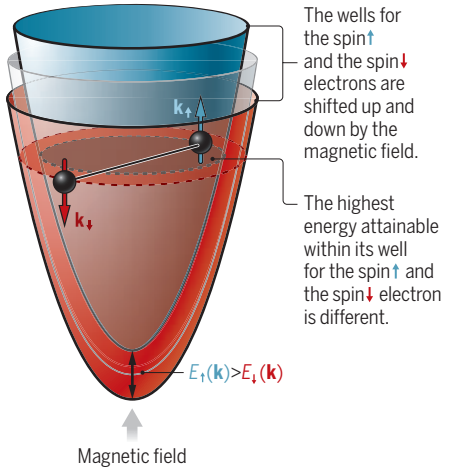


The well represents the range of possible energies (E) for the electrons with the momentum ($\mathbf{k}$) inside the material.

The highest energy level is also the energy of the electrons forming the pair.

Fulde-Ferrell-Larkin-Ovchinnikov electron pair

Momentum: $\mathbf{k}_\uparrow + \mathbf{k}_\downarrow \neq 0$



The wells for the spin↑ and the spin↓ electrons are shifted up and down by the magnetic field.

The highest energy attainable within its well for the spin↑ and the spin↓ electron is different.

GRAPHIC: V. ALTOUNIAN/*SCIENCE*

Institute for Advanced Simulation, Forschungszentrum Jülich, 52425 Jülich, Germany. Email: e.pavarini@fz-juelich.de

evidence against the chiral spin-triplet picture was collected (*13*–*15*). This paradigm change also implied that the Zeeman splitting could indeed break pairs, and thus the stage could be ideal for an FFLO state.

Kinjo *et al.* found signatures of an FFLO state, a result in line with the dismissal of the chiral spin-triplet picture. By measuring the nuclear magnetic resonance (NMR) signal at specific sites of the crystal lattice, they measured the spatial variation in spin density at 70 mK—well below the superconducting critical temperature. The authors gradually increased the magnetic field while making sure that the entire volume of the sample remained in the superconducting state up until the critical field of 1.4 T. At low magnetic fields, Sr_2RuO_4 behaved like a conventional superconductor, with a single NMR peak accompanied by a resonance frequency shift, which is known as Knight shift. This shift was smaller in the superconducting phase than in the normal phase, which is in line with the expectations for conventional superconductivity. However, approaching the critical magnetic field, a second peak appeared in the NMR signal. The associated Knight shift was larger than in the normal phase, which is something that cannot be easily explained by a mixture of normal and superconducting phases and instead could be consistent with spatial spin-density modulations associated with an FFLO state.

The discovery of a possible FFLO state in the layered perovskite Sr_2RuO_4 paves the way for unprecedented studies of the elusive superconducting state. The evidence presented by Kinjo *et al.* is compelling but still indirect. The smoking gun—the direct measurement of the spatial modulations of the superconducting order parameter—remains at large. For now, Sr_2RuO_4 confirms itself as a system of neverending wonders. It might not be a chiral spin-triplet superconductor, but it is certainly one of a kind. More surprises could be in store for the coming years. ■

REFERENCES AND NOTES

1. J. Bardeen *et al.*, *Phys. Rev.* **108**, 1175 (1957).
2. P. Fulde, R. A. Ferrell, *Phys. Rev. B* **135**, A550 (1964).
3. A. I. Larkin *et al.*, *Zh. Exp. Teor. Fiz.* **47**, 1136 (1964).
4. K. Kinjo *et al.*, *Science* **376**, 397 (2022).
5. H. A. Radovan *et al.*, *Nature* **425**, 51 (2003).
6. A. Bianchi *et al.*, *Phys. Rev. Lett.* **91**, 187004 (2003).
7. K. Cho *et al.*, *Phys. Rev. B* **79**, 220507R (2009).
8. H. Mayaffre *et al.*, *Nat. Phys.* **10**, 928 (2014).
9. G. Zhang *et al.*, *Phys. Rev. Lett.* **116**, 106402 (2016)
10. T. M. Rice, M. Sigrist, *J. Phys. Condens. Matter* **7**, L643 (1995).
11. K. Ishida *et al.*, *Nature* **396**, 658 (1998).
12. S. Yonezawa, T. Kajikawa, Y. Maeno, *Phys. Rev. Lett.* **110**, 07703 (2013).
13. A. Pustogov *et al.*, *Nature* **574**, 72 (2019).
14. K. Ishida *et al.*, *J. Phys. Soc. Jpn.* **89**, 034712 (2020).
15. A. N. Petsch *et al.*, *Phys. Rev. Lett.* **125**, 217004 (2020).

10.1126/science.abn3794

CANCER

A fresh look at somatic mutations in cancer

Analysis of cancer genome sequences reveals new mutational signatures

By **Dávid Szüts**

Sequencing the genomic DNA of tumor samples has provided rich data on mutagenic processes and revealed that information derived from heterogeneous mutational landscapes of cancers can complement the identification of driver gene mutations in aiding reliable personalized diagnosis and treatment selection. On page 368 of this issue, Degasperi *et al.* (*1*) present a comprehensive analysis of the patterns of somatic base substitution mutations in human cancers. Their work builds on landmark studies that revealed that mutational catalogs, or spectra, can yield constituent mutational components that reflect biological processes (*2*–*4*). Combining component "signatures" can adequately reconstitute the mutation spectrum of individual cancer samples and facilitate an understanding of the underlying mutagenic mechanisms.

Mutational signatures categorize mutations according to the sequence context and the type of change, which are generally specific to an external mutagen or endogenous mutagenic process. There are two fundamental approaches in mutational signature analysis: using unsupervised algorithms to define de novo signatures in a dataset and deconstructing mutation spectra into predefined signatures. Human cancers are expected to display a finite variety of mutational processes, which fuels the endeavor to establish a comprehensive set of cancer-associated mutational signatures, and their potential clinical use would benefit from a scientific consensus. Research in this area has progressed rapidly, and by 2021, the Catalogue of Somatic Mutations in Cancer (COSMIC) listed ~60 real single-base substitution (SBS) signatures and several artifactual ones (*5*). The increasing number of signatures, which were not all derived from a single analysis, have made it progressively more complex to fit signatures to new datasets, with a judicious preselection of mutational signatures often necessary for the most meaningful results.

Institute of Enzymology, Research Centre for Natural Sciences, Budapest, Hungary. Email: szuts.david@ttk.hu

The study by Degasperi *et al.* used an elegant and clear solution for defining mutational signatures, initiated using organ-specific sample cohorts. A stable two-step process first found common signatures in spectra from samples that cluster with others by similarity, and then the spectra from excluded samples were deconstructed using defined common signatures and potential additional rare signatures. The organ-specific signatures were then clustered, filtered, and averaged to establish 120 reference signatures that present a picture of general mutational processes in cancer (see the figure). The palette of common signatures appears to be saturated, but the approach leaves the door open to finding rare signatures.

The greatest asset of the study is the notable 12,222-sample dataset from the 100,000 Genomes Project of Genomics England, which has a greater number of whole-genome sequences than previous major cancer sequencing projects put together (*6*). The results were validated on datasets from the International Cancer Genome Consortium and the Hartwig Foundation. The Genomics England dataset contains hundreds of samples per organ type, which allowed the detection of rare signatures that were present in less than 1% of each organ set.

What has the new analysis revealed? Beyond confirming most, but not all, of the previously identified base substitution signatures, it added 40 new SBS and 18 double-base substitution signatures to the reference set. The signatures bear evidence of the expected mutagenic processes: effects of cellular metabolism, deficiencies of DNA repair, and environmental sources. Of particular interest are signatures that only show subtle differences but can be tied to different mechanisms through association with driver gene mutations, such as inactivating methyl-CpG binding domain protein 4 (*MBD4*) mutations that explain signature SBS96, which is similar to other signatures caused by CpG deamination (*7*). Because the mechanism behind many signatures is still unknown, the

study of Degasperi *et al.* also provides fertile ground for further investigations.

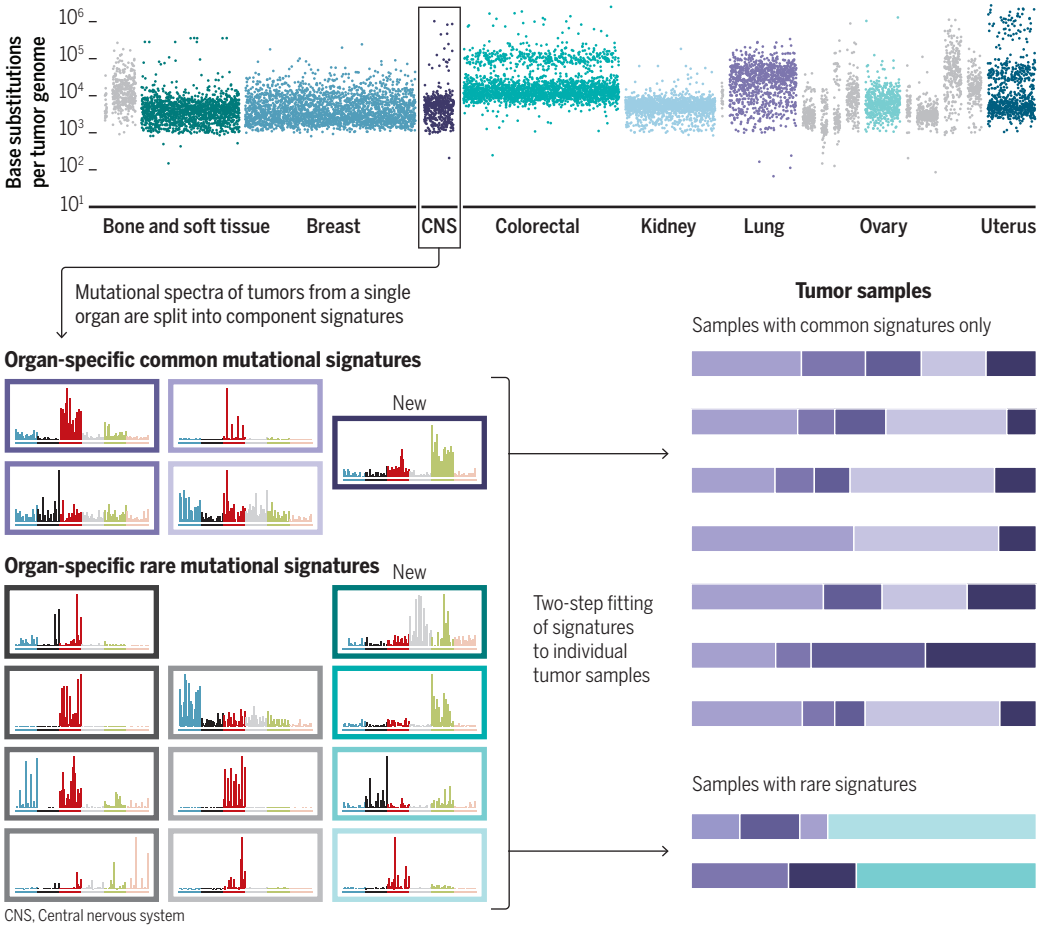
The mechanisms that lead to mutational signatures are most often validated or elucidated in cultured cells, using environmental mutagenic treatments or genetically modified cell lines (*8*–*10*). Such experiments can provide clear mutation spectra, but causative connections must be drawn carefully because different processes may produce very similar spectra. For a better understanding, investigators may also focus on attributes beyond the immediate sequence context, such as transcription or replication strand bias (which were also reported by Degasperi *et al.*), clustering, associations with replication timing, wider sequence context, and chromatin features. Such information can be handled separately or incorporated into signatures (*11*), with the latter approach complicated by multidimensionality, or a need to assign relative weights to the components. The number of mutations in individual samples also limits the complexity of analyses.

The primary incentive of cancer mutational signature analysis is to provide diagnostic tools for patient stratification (*12*). The first requirement is the fitting of signatures to mutational spectra of individual samples. Degasperi *et al.* provide a method that is analogous to signature decomposition: Common signatures are fitted first, followed by the addition of rare signatures one at a time. Fitting organ-specific signatures defined on the largest datasets is likely to be particularly accurate, and it may also help determine the organ of origin of metastatic samples. The scale of the analysis demonstrated that multiple organs have rare cases of tumors with signatures that are common in other organs, for example, those attributed to deficiency of DNA repair through homologous recombination or mismatch repair, which may benefit from targeted treatment.

Reliable information is needed on the predictive value of mutational signatures for drug sensitivities, which is most likely to find use in the treatment of DNA repair–deficient tumors. Many samples with DNA repair deficiency–associated signatures lack identifiable driver gene mutations because of an incomplete understanding of the underlying biology or chromatin-dependent gene expression changes, arguing for the preferred use of mutational signature analysis rather than targeted gene sequencing. Indeed, a preliminary study reports that the association of drug sensitivities with mutational signatures is stronger than that with driver mutations (*13*). However, both in vitro and clinical investigations have found that the expected correlation between mutational signatures connected with homologous recombination deficiency and sensitivity to poly(ADP-ribose) polymerase (PARP) inhibitors (which target this defect) is less than straightforward (*14*, *15*). The use of compound classifiers that rely on multiple mutation types may help provide more accurate predictions. It must also be remembered that mutational spectra give a compressed view of the tumor's past, and the observed mutational processes may no longer operate at the time of treatment. ■

Mutational signature analysis in cancer

Whole-genome sequences of 12,222 tumor-normal sample pairs from the 100,000 Genomes Project of Genomics England reveal single-base substitutions in different cancer types (top). These mutation spectra were analyzed by Degasperi *et al.* to define component mutational signatures (bottom), which are induced by specific mechanisms (many of which are unknown). These mutational signatures give information about the history of mutagenesis and properties of individual tumors.



CNS, Central nervous system

GRAPHIC: K. FRANKLIN/*SCIENCE*

REFERENCES AND NOTES

1. A. Degasperi *et al.*, *Science* **376**, eabl9283 (2022).
2. S. Nik-Zainal *et al.*, *Cell* **149**, 979 (2012).
3. L. B. Alexandrov *et al.*, *Nature* **500**, 415 (2013).
4. L. B. Alexandrov *et al.*, *Nature* **578**, 94 (2020).
5. COSMIC, Mutational signatures (v3.2 - March 2021); http://cancer.sanger.ac.uk/cosmic/signatures.
6. C. Turnbull, *Ann. Oncol.* **29**, 784 (2018).
7. M. A. Sanders *et al.*, *Blood* **132**, 1526 (2018).
8. J. E. Kucab *et al.*, *Cell* **177**, 821 (2019).
9. X. Zou *et al.*, *Nat. Commun.* **9**, 1744 (2018).
10. J. Zámborszky *et al.*, *Oncogene* **36**, 746 (2017).
11. H. Vöhringer *et al.*, *Nat. Commun.* **12**, 3628 (2021).
12. S. W. Brady *et al.*, *Trends Genet.* **38**, 194 (2022).
13. J. Levatić *et al.*, *bioRxiv* 2021.2005.2019.444811 (2021).
14. Á. Póti *et al.*, *Genome Biol.* **20**, 240 (2019).
15. N. Chopra *et al.*, *Nat. Commun.* **11**, 2662 (2020).

ACKNOWLEDGMENTS

I thank E. Németh for help with the figure.

10.1126/science.abo7425

GENETICS

Repetitive DNA in disease

Transposons become a focus of speculation and scrutiny in biomedical research

By **Kathleen H. Burns**[1,2]

In 1988, physician-scientists studying the genetic basis of hemophilia discovered pathogenic long interspersed element 1 (LINE-1) insertions in the coagulation factor VIII gene (*1*). Therein was unequivocal evidence of a new mutagenic mechanism in humans: insertions of mobile genetic elements (transposons) originating from the mostly silent, repetitive sequences of the human genome. In subsequent years, insertional mutagenesis proved a recurrent, albeit uncommon, cause of genetic disease. Now, a paradigm shift is placing new emphasis on transposon control as a vital cellular function that is compromised by and potentially contributes to many diseases. Transposon expression and activity is perhaps most overt in cancers, although aberrant expression of LINE-1, endogenous retroviruses (ERVs), and other repeats have also been invoked as a dimension of developmental, degenerative, and autoimmune diseases. As rigorous and accessible tools are developed to investigate, new discoveries will clarify the roles of repetitive elements and their regulation in disease.

The human genome, like other complex eukaryotic genomes, is filled with highly repetitive DNA attributable to the activity of self-propagating genetic sequences. In humans, this landscape is dominated by retrotransposons that make new copies of themselves by first being transcribed to RNA intermediates and then reverse transcribed to cDNA that is ultimately integrated into the genome. Altogether, nearly half of the human nuclear genome is transposon-derived—a record of the activity of various types of retrotransposons over millennia. The overwhelming majority of these sequences are incomplete or mutated and no longer active as mobile elements; however, there are exceptions. Subfamilies of LINE-1 retain the ability to encode proteins that make new genomic copies of LINE-1 and other noncoding repeats—namely *Alu* short interspersed elements (SINEs) and composite SINE-variable number tandem repeat (VNTR)–*Alu* (SVA) elements. Other functions can reside in transposable elements that are now inactive as insertional mutagens. For example, many ERVs [long terminal repeat (LTR) sequences] contain strong promoters or enhancers that can act on nearby genes. To varying degrees, these elements retain residual protein-coding capacity as well.

LINE-1 sequences make copies of themselves by producing a 6-kilobase (kb), bicistronic RNA intermediate transcribed from an internal RNA polymerase II promoter. The first of its open reading frames (ORFs) encodes ORF1p, which forms an RNA-binding homotrimer. The second ORF encodes ORF2p, which encompasses endonuclease and reverse transcriptase domains, which are responsible for making cDNA copies of the LINE-1 RNA and inserting these into the genome. ORF2p functions can also be co-opted by noncoding *Alu* and SVA elements to reverse transcribe and integrate their sequences into the genome instead of the protein-coding LINE-1 RNA. Collectively, LINE-1, *Alu*, and SVA insertions are uncommon but recurrent causes of genetic mutation, leading to various constitutional genetic diseases. In almost every case, the transposable element inserts into a critical portion of a gene and causes a loss-of-function allele.

Disease-causing insertions can occur in or be transmitted by the germline and thus establish an allele frequency in a human population. Another effect of germline transmission is that each individual inherits a different complement of active transposons (*2*). It may follow that the expression of mobile elements and the burden of retrotransposition consequently varies from person to person, and whether there are limits to what is tolerable is unknown. During development and in adults, transposition may contribute to genetic heterogeneity in various tissues (somatic mosaicism), but to what degree and whether this is pathogenic are difficult to know. Baseline activity may be physiologic. It has been hypothesized that somatic retrotransposition is a source of functional genetic diversity in normal neuronal development. However, single-cell sequencing experiments indicate that the level of somatically acquired insertions in neurons is very low, making a required role for these insertions in normal development seem unlikely (*3*).

Like gene expression, transposable element expression patterns are cell type–specific and can depart from normal in diseased states. For example, LINE-1 promoter hypomethylation and LINE-1 overexpression are hallmarks of many cancers, and ORF1p accumulates in several of the most common and commonly lethal human malignancies (*4*) in association with loss of the *TP53* tumor suppressor gene (*5*). Other types of retroelements show altered expression in cancers as well. This inspires practical questions, such as whether these can be used as biomarkers of disease—as well as more fundamental questions, such as what regulatory mechanisms are responsible for expression of transposable elements, and whether the expression of transposable elements is epiphenomenal or causes malignancy.

Transposable element–encoded regulatory sequences can alter chromatin states and affect nearby gene expression (see the figure). Thus, transposon activation may underlie or be coordinated with gene expression changes that promote transformation, although demonstrations of this are limited. What about transposon activation independent of these regulatory effects? There are anecdotal cases in which an individual has inherited a LINE-1 locus that escaped silencing by promoter methylation and went on to "drive" tumor development through insertional mutagenesis, disrupting expression of the adenomatous polyposis coli (*APC*) tumor suppressor gene in colonic epithelium and leading to colon cancer (*6*). Overt cancer-driving mutations such as these are rare, however.

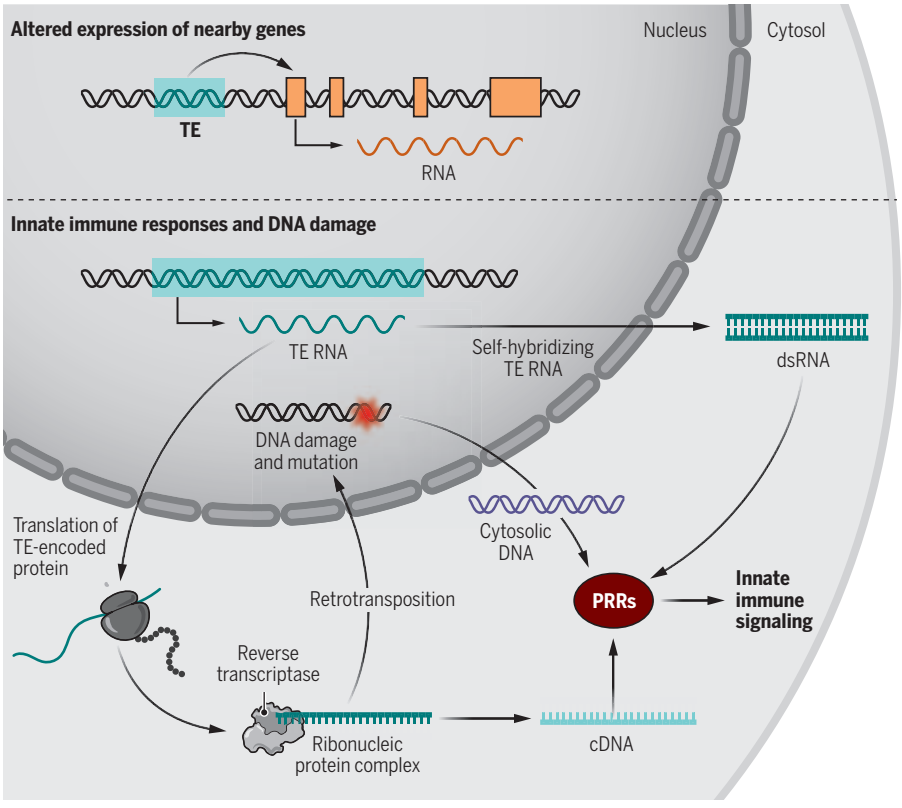
Somatically acquired LINE-1 insertions that appear to be "passengers" (nondriver alterations) in cancer genomes are much more commonplace (*7*). If individual insertions are functionally unimportant, do they collectively represent a meaningful mutational signature? Recent studies indicate that LINE-1 insertion intermediates can cause chromosomal rearrangements (*7*) and pose barriers to DNA replication (*8*). These data suggest that DNA lesions induced during transposition may have underrecognized roles in shaping cancer genomes and that acquiring tolerance for LINE-1 activity may be a prerequisite for tumorigenesis.

Given the widespread expression of LINE-1 in cancers, it may be an intuitive assumption that transposable element activation promotes malignant growth. Paradoxically, however, imposing transposon expression in experimental systems is often cytotoxic. Expression of LINE-1, *Alu*, and ERV elements can induce in-

[1]Dana Farber Cancer Institute, Boston, MA, USA.
[2]Harvard Medical School, Boston MA, USA.
Email: kathleenh_burns@dfci.harvard.edu

Cellular effects of transposable element dysregulation

Transposable element (TE) activation can directly regulate gene expression when the locus encodes a regional promoter or enhancer (top). TE dysregulation can also have cellular effects (bottom); for example, innate immune pattern recognition receptors (PRRs) can be activated by DNA damage or cDNA produced by retroelement proteins, or by self-hybridizing double-stranded RNA (dsRNA).



nate immune responses, triggering nucleic acid–sensing pattern recognition receptors and interferon signaling, much like infection by an exogenous virus. Depending on the type of transposon and cell context, these innate immune responses may be mediated by DNA damage, by the activity of retrotransposon-encoded reverse transcriptase, or by the presence of self-hybridizing double-stranded RNAs (dsRNAs). The resulting so-called "viral mimicry" can curtail tumor growth through cell autonomous mechanisms (inducing cell cycle arrest or limiting protein translation) or by prompting clearance of the cells by the immune system. Thus, in the early stages of transformation, epigenetic activation of transposons may prevent or restrain tumor development (*9*). In fully developed cancers, transposon activity may promote genome evolution and the emergence of tumor cell clones with particular characteristics. Additionally, malignancies could be at a "tipping point," at which enhancing transposition is unsustainable or limits tumor growth under particular conditions. Epigenetic agents to enhance transposon expression, for example, may enhance responses to immunotherapies or create molecular dependencies (*10*–*12*). To the extent that transposons can be leveraged to incite DNA damage, this strategy may sensitize cancers to genotoxic agents or to DNA repair inhibitors (*8*).

Beyond cancer, transposon-induced innate immune signaling or DNA damage is implicated in aging and various diseases. In aging, reduced expression of the 3′ exonuclease TREX1 (three-prime repair exonuclease 1) and up-regulated activity of LINE-1 ORF2p together may promote accumulation of cytoplasmic cDNA, inducing senescence-associated interferon signaling (*13*). Mutations compromising *TREX1* in Aicardi-Goutières syndrome (AGS) may have an analogous interaction with transposons that fuels pathologic interferon signaling and encephalopathy (*14*). This hypothesis led to ongoing clinical trials of reverse transcriptase inhibitors in AGS. Reverse transcriptases are one of several protein types encoded by the "noncoding" genome, and others may be relevant to disease. For example, ERV expression has been associated with the neurodegenerative disease amyotrophic lateral sclerosis, and expression of an ERV-encoded env protein in mice causes DNA damage and motor neuron degeneration through unknown mechanisms (*15*). The hypothesis that retroelement expression is pathogenic is now being pursued for several neurodegenerative diseases.

There are a host of technical needs to better understand the role of transposons in disease. Interspersed repeats are variable and exist in high copy numbers in the genome, and they are pervasively incorporated in pre-mRNAs and long noncoding RNAs (lncRNAs). Therefore, finding DNA insertion variants or differentially expressed loci in sequencing datasets is challenging. Reference genomes and genotyping tools are needed that capture heritable variants in active transposons. Single-cell sequencing approaches are needed to indicate de novo insertions and associated genome rearrangements, to reveal expression of specific transposon loci, and to profile repeat-containing cDNA and dsRNA species. Reagents to study this biology in diseased tissues and in clinical specimens will also be important. Most nucleic acid probes and primers designed against a repetitive sequence will indiscriminately hybridize with unintended targets that exist throughout the genome. If overexpression of a transposable element is technically straightforward to recreate in experimental models and sufficient to cause disease, tools are needed to ascertain whether it is a necessary contributor. Removing retroelements from a disease model is challenging, especially because they are often species-specific and exist among many highly homologous sequences. Models that use human tissue, locus-specific genetic approaches, and pharmacologic agents that precisely target retroelements are needed for functional studies and to set the stage for our bench-to-bedside aspirations. ■

REFERENCES AND NOTES

1. H. H. Kazazian Jr. *et al.*, *Nature* **332**, 164 (1988).
2. C. R. Beck *et al.*, *Cell* **141**, 1159 (2010).
3. G. D. Evrony *et al.*, *eLife* **5**, e12966 (2016).
4. N. Rodić *et al.*, *Am. J. Pathol.* **184**, 1280 (2014).
5. B. Tiwari *et al.*, *Genes Dev.* **34**, 1439 (2020).
6. E. C. Scott *et al.*, *Genome Res.* **26**, 745 (2016).
7. B. Rodriguez-Martin *et al.*, *Nat. Genet.* **52**, 306 (2020).
8. D. Ardeljan *et al.*, *Nat. Struct. Mol. Biol.* **27**, 168 (2020).
9. Y. Zhao *et al.*, *Nat. Immunol.* **22**, 1219 (2021).
10. D. Roulois *et al.*, *Cell* **162**, 961 (2015).
11. K. B. Chiappinelli *et al.*, *Cell* **162**, 974 (2015).
12. P. Mehdipour *et al.*, *Nature* **588**, 169 (2020).
13. M. De Cecco *et al.*, *Nature* **566**, 73 (2019).
14. C. A. Thomas *et al.*, *Cell Stem Cell* **21**, 319 (2017).
15. W. Li *et al.*, *Sci. Transl. Med.* **7**, 307ra153 (2015).

ACKNOWLEDGMENTS

K.H.B. is supported by the National Institutes of Health (R01GM130680 and R01CA240816). She is a Scientific Advisory Board member of Transposon Therapeutics and the scientific cofounder of Oncolinea Pharmaceuticals.

10.1126/science.abl7399

GRAPHIC: KELLIE HOLOSKI/*SCIENCE*

POLICY FORUM

SCIENCE AND SECURITY

Countering the future chemical weapons threat

Shift focus from disarmament to preventing reemergence

By **Tuan H. Nguyen**

After decades of difficult negotiations, the Chemical Weapons Convention (CWC) was adopted in 1993 and entered into force on 29 April 1997, banning the development, production, stockpiling, transfer, and use of chemical weapons (CW). As the CWC celebrates the 25th anniversary of its entry into force, it can document considerable success, much of it attributed to the CWC implementing body—the Organisation for the Prohibition of Chemical Weapons (OPCW). Yet, facing a volatile international security environment and an everchanging chemical industry, the OPCW must transform to meet its mission and remain an exemplar for multilateralism. As the next CWC review conference approaches in 2023, a next-generation OPCW 2.0 can be effective and credible only if it reinforces international norms against CW, anticipates future challenges posed by advancements in science and technology (S&T), incorporates more qualitative elements into the verification and compliance system, and keeps pace with technological change.

The CWC is the first multilateral disarmament agreement to provide for the comprehensive ban, including elimination, of an entire category of weapons of mass destruction (WMD). The treaty has almost universal membership. The OPCW, which was awarded the Nobel Peace Prize in 2013, has conducted more than 4200 total industry inspections and overseen the destruction of 99% (71,614 metric tons) of the world's declared CW stockpiles (*1*).

TOWARD THE OPCW 2.0

Since entry into force, the OPCW's primary focus has been on the elimination of existing CW stockpiles and production capacities and the prevention of their acquisition in the future. The OPCW routine verification activities center around the continuous monitoring of CW stockpile destruction and the highly intrusive on-site inspections of industry facilities to ensure that they are not diverted for CW-related activities. By September 2023, all remaining declared CW stockpiles and equipment are to be eliminated (*2*). The OPCW will transition from an organization focused on chemical disarmament to one dominated by nonproliferation and threat-reduction activities. Given the recent use of CW in a civil war and for assassinations, the OPCW will increasingly focus on states and nonstate actors that are developing or using CW.

Thus, the OPCW 2.0 will see time and resources that are now devoted to the elimination of declared stocks repurposed for nonroutine verification activities. These include investigations of alleged CW use, challenge inspections, and more routine industry inspections, especially of other chemical production facilities (OCPFs), which manufacture large quantities of discrete organic chemicals that contain phosphorus, sulfur, or fluorine—the common constituents of blister and nerve agents—that are not on the CWC schedule of banned and controlled chemicals. To remain effective and credible, the challenge for the OPCW is that it must transform itself as a technical organization in a highly polarized political environment.

REINFORCING NORMS

For its survival and relevancy in the future, negotiators recognized the need for the CWC to be a dynamic document that empowers the OPCW to evolve, adapt, and adjust for the possibility of improvement based on a changing threat environment. The OPCW 2.0 must reinforce international norms against CW by exercising existing verification mechanisms and strengthening them by enhancing its technical capacity and developing new technical tools. Many questions arise. How do the States Parties and OPCW reduce the benefit and utility of CW programs to violators? Should the States Parties and OPCW seek to invoke the challenge inspection mechanism, which is intended as a verification safety net to capture and deter clandestine and undeclared CW activities located in another member state?

Case studies can shape answers to such questions. Knowledge gained from the United Nations (UN) Special Commission's experience in dealing with CW stocks in Iraq after the Iran-Iraq War and Desert Storm in the 1980s and early 1990s, respectively, informed the negotiation of the CWC and the creation of the OPCW. This experience was critical to developing the intrusive industry verification regime, which is the hallmark of the CWC and the reason why it can still be effective in the future. More recently, the mission in Syria was a watershed moment for the OPCW. The coordination and cooperation by the international community in 2014 to remove, transport, and destroy Syria's CW stockpile was a powerful demonstration of effective multilateralism. However, the OPCW has much more work ahead. Allegations of CW use by Syria on multiple occasions from 2014 to 2018, as well as conclusions by the OPCW Fact-Finding Mission that CW were used on multiple occasions and by the OPCW-UN Joint Investigative Mechanism that attributed the sarin and chlorine attacks to the Syrian regime and the sulfur mustard attacks to the Islamic State of Iraq and the Levant (ISIL), are still denied by the perpetrators and their partners.

The standard-of-proof issue has been a long-standing problem for verification and compliance of arms control and disarmament agreements. It is a tall order to find, identify, and attribute activities that provide undisputable evidence of treaty violation. Without the ability to hold violators accountable, the CWC and OPCW will decrease in effectiveness in a postdisarmament future after all declared CW stockpiles have been destroyed. A step in the right direction was the 2018 decision by the States Parties granting the OPCW the mandate to investigate and attribute responsibility of alleged CW use in Syria, which included the establishment of an Investigation and Identification Team that functions under the direct authority of the OPCW director-general (*3*).

In addition to the mission in Syria, the OPCW has assisted in the investigation and confirmation of other recent incidents of alleged CW use. The OPCW provided technical assistance to the Malaysian government in 2017 when the VX nerve agent was used to assassinate Kim Jong-nam, the half-brother of Kim Jong-un, the leader of North Korea. Technical assistance from the OPCW was provided again in 2018 when a novichok agent was used in the assassination attempt on Sergei and Yulia Skripal in Salisbury, UK. Novichoks are a class of organophosphorus nerve agents developed by the former Soviet Union during the Cold War that were not listed on the

Center for Global Security Research and Global Security Directorate, Lawrence Livermore National Laboratory, Livermore, CA, USA. Email: thn@llnl.gov

CWC Annex on Chemicals at the time of the Salisbury attack. The OPCW confirmed the UK's finding that a novichok was used in the attack but did not provide attribution identifying the perpetrators of the assassination attempt. Nevertheless, the Salisbury case resulted in novichoks being added to Schedule 1 of the CWC Annex on Chemicals—the first time that the annex has been updated. Despite this, a novichok-type agent was used again in the 2020 assassination attempt on Alexei Navalny (*4*).

The lack of accountability for these incidents erodes and weakens the international norm against CW. The use of CW perpetuates doubts about the compliance of some States Parties and undermines the credibility and overall effectiveness of the CWC. The international community must be assured that all CW programs declared by member states are completely dismantled. From a disarmament perspective, this is the fundamental reason why accountability and credible attribution of CW use matter.

To address and investigate claims of noncompliance, negotiators included provisions in the CWC for member state–initiated mechanisms for the OPCW to investigate allegations of CW use and to conduct short-notice challenge inspections. To date, no member state of the CWC has requested a challenge inspection. There are many reasons for this, both technical and political. The standard-of-proof challenge raises doubts about the OPCW's technical capacity to find, identify, and attribute violations during a challenge inspection. The political risks of failure during a challenge inspection to find indisputable evidence to substantiate a cheating or violations charge could end up provoking a cycle of retaliatory challenge inspections. As such, challenge inspections should be thought of not only as a mechanism for cases where there is clear evidence of cheating and violations but also as an instrument to clarify ambiguities, resolve declarations inconsistencies and discrepancies, and address concerns about compliance to prevent violators from circumventing the CWC and undermining its effectiveness.

If the challenge inspection mechanism continues to remain unused, its deterrence value will decline. To aid in these investigative efforts, the OPCW 2.0 needs to enhance and further develop technical competencies in areas that include forensic science, evidence collection, data management, and crime-scene reconstruction. The 2019 report by an OPCW working group on investigative S&T provides activities that the OPCW should consider to enhance its technical and operational capabilities and capacity to provide technical assistance to member states and to effectively conduct nonroutine missions, such as fact finding and investigations for attribution (*5*).

ANTICIPATING S&T CHALLENGES

Advances in S&T lower the barriers for WMD proliferation by enabling alternative proliferation pathways that present challenges for verification and compliance regimes. The convergence of knowledge in the life and physical sciences and engineering disciplines is enabled and accelerated by information technology and expanded access to research findings, data, and advanced computing and algorithms. Advances in gene editing, biocomputations, and predictive biology tools coupled with systems biology will enable the design and tailoring of molecules that are well suited for an intended role or purpose. New additive manufacturing techniques, along with other just-in-time production technologies, can challenge verification and compliance protocols.

Technicians with the Organisation for the Prohibition of Chemical Weapons conduct a mock investigation of dummy chemical weapons in Rijswijk, the Netherlands, in 2017.

Yet many dual-use technologies required for CW production have been known for more than 70 years. The production of organophosphorus nerve agents is not fundamentally different from the production of commercial pesticides. Still, the synthesis of some organophosphorus nerve agents requires a few chemical reactions that are not common in the production of commercial pesticides, such as the fluorination and alkylation reactions that form the phosphoryl-fluoride functional group and phosphoryl-carbon bonds. The phosphoryl-carbon bond is critical to the physicochemical properties of organophosphorus nerve agents and guided the organization of the schedules of the CWC Annex on Chemicals.

To achieve phosphoryl-carbon bonds by the Michaelis-Arbuzov rearrangement reaction, one of the most important and fundamental reactions in organophosphorus chemistry, heat must be applied under traditional conditions. Large-scale production requires industrial-size heating mantles, heat exchangers, other specialized equipment, and safety containment measures. It had long been assumed that CW-capable facilities would have large footprints and distinctive signatures and be limited in number. But modern synthetic organic chemistry provides alternative methods and pathways. This same rearrangement reaction to form phosphoryl-carbon bonds can occur without heat and instead be catalyzed by visible light or common organic reagents at room temperature (*6*–*8*). Microwave energy can be used to efficiently carry out the transformation without the need for traditional heat sources (*9*). Most future WMD threats may thus have smaller footprints and less distinct signatures and be more closely associated with legitimate civilian-industrial activities rather than with highly centralized, large-scale weapons programs run by military entities (*10*).

The renaissance of photochemical and electrochemical methods for chemical synthesis is bringing forth a multitude of innovative chemical transformations. For decades, some of the prevailing challenges facing photochemistry and electrochemistry have been reproducibility, scale, and efficiency. Many of these issues may now be overcome by using continuous-flow platforms developed in collaboration with engineers and computer scientists. The scale-up potential of this approach is demonstrated by the synthetic output of a desired product at a rate exceeding 5 kg/day from simple precursors (*11*). The Michaelis-Arbuzov reaction can now be carried out in a continuous-flow and solvent-free method (*12*). Photocatalytic fluorinations have also been demonstrated in continuous flow as a mild alternative to traditional methods that are both highly toxic and corrosive, avoiding

PHOTO: OPCW/CC BY-NC 2.0

equipment lined with corrosion-resistant alloys and other specialized materials (*13*).

Such changes can make possible what was once thought challenging or impossible. Combined with changes in process chemistry enabled by automation, robotics, and algorithms to assist with synthetic design, reaction prediction, and starting material selection, modern synthetic organic chemistry will challenge the way verification inspections are done. Will inspectors be looking for the right signatures for CW production? Or should they be thinking about nontraditional precursors, nontraditional equipment and facilities, and maybe even nontraditional people? Mechanisms for verification and compliance of these new developments are not as clear, even though the CWC, more than any other multilateral arms control treaty, is highly detailed with its comprehensive verification system. The challenge for the OPCW 2.0 will be to ensure that inspectors are aware of these platform technologies, recognize their enabling capabilities, and understand the potential for CW production.

BECOMING MORE QUALITATIVE

Lessons can be learned from the Nuclear Non-Proliferation Treaty's 1997 Additional Protocol, which was a direct response to the discovery of secret nuclear weapons programs in Iraq and North Korea in the early 1990s (*14*). The Additional Protocol sought to strengthen and expand existing International Atomic Energy Agency (IAEA) safeguards for verification of nuclear technology. The purpose was to provide the international community with a measure of confidence that any diversion of material and technologies to a weapons program would be detected quickly.

To accomplish this, the underlying notion of the Additional Protocol was to restructure the IAEA safeguards regime from a quantitative system to one that was more qualitative. Instead of focusing only on accounting for declared quantities of nuclear material and monitoring declared activities, the Additional Protocol sought to focus more on providing a comprehensive picture of a State Party's activities. To achieve this, the amount and type of information provided to the IAEA expanded, the number and types of facilities subject to inspection increased, and environmental sampling during such inspections was made possible.

Today, the CWC verification regime is mostly quantitatively focused on complete accountability of all items declared for CW programs, whereas industry verification is basically to confirm the nonproduction of CW. OPCW inspectors do not have the authorization to detect chemicals other than those listed in the official OPCW analytical database, which is limited to scheduled chemicals. Although an Additional Protocol for the CWC is not appropriate because the nuclear and chemical threats are different, challenges posed by advances in S&T require the OPCW 2.0 to evolve the industry verification paradigm to become more qualitative to provide a more comprehensive picture of activities at declared facilities. For example, the site-selection methodology for inspection of OCPFs needs to be refined to target facilities that pose the greatest risk for diversion to CW production—perhaps inclusion of open-source information to enhance the verification process.

KEEPING PACE WITH TECHNOLOGY

In the past, countering the CW threat was primarily focused on keeping precursor materials and/or toxic chemicals out of the hands of proliferators by securing knowledge, specialized materials, and dedicated equipment. That strategy was believed to be effective because the infrastructure to support large-scale military CW campaigns was expected to have a large footprint, have many signatures, and require considerable scale-up time. While looking for large, questionable chemical facilities, efforts focused on regulation and control of construction or source-material transport. But today, substantial amounts of materials and knowledge can be gathered outside traditional supply chains, including through the internet. Proliferators can find recipes for chemical agents, as exemplified by the reemergence of illicit synthetic designer-drug networks that seem to stay one step ahead of regulators at a time when we face the prospect that substantial amounts of CW could be produced in low-profile facilities. New chemical production processes and technologies enable just-in-time production, the widespread adoption of which could facilitate treaty breakout scenarios in which a member state vacates disarmament commitments, reneges on treaty obligations, and rapidly rearms (*15*).

Fortunately, many of the same breakthrough technologies that will challenge the OPCW 2.0 will also afford the organization new enabling capabilities to accomplish its mission more effectively. Instead of relying solely on data from site inspections and declared information, more comprehensive assessments can be made with data analytic tools being adopted in the commercial sector, including data aggregation tools, space-based assets for "eyes on the ground" capabilities, and open-source information products. Advances in remote sensing and automated sampling systems could transform the way we think about and execute monitoring and verification activities. The OPCW 2.0 needs to keep pace with technological developments to improve the conduct of verification by adopting new methods and types of inspection equipment.

The mission of the OPCW 2.0 will not change: to implement the provisions of the CWC to advance the vision of a world free of threats from CW. The issue is not about what to do but how to do it. The greatest challenge for the OPCW 2.0 is not in the legal framework of the CWC or the tools needed to adapt, but rather the political will to use them. After 25 years, the CWC and the OPCW are at a crossroads. The path that the OPCW 2.0 takes will determine whether the CWC continues to be celebrated as a model for multilaterally negotiated arms control and disarmament and a central pillar to our strategy to counter WMD threats or risks becoming less effective in the future international security environment. ■

REFERENCES AND NOTES

1. OPCW, OPCW by the numbers; https://www.opcw.org/media-centre/opcw-numbers [accessed 26 March 2022].
2. Program Executive Office of the Assembled Chemical Weapons Alternatives (ACWA), U.S. chemical weapons stockpile destruction progress; https://www.peoacwa.army.mil/destruction-progress/ [accessed 26 March 2022].
3. OPCW, "Decision: Addressing the threat from chemical weapons use" (C-SS-4/DEC.3, OPCW, 2018); https://www.opcw.org/sites/default/files/documents/CSP/C-SS-4/en/css4dec3_e_.doc.pdf.
4. OPCW, "OPCW issues report on technical assistance requested by Germany," 6 October 2020; https://www.opcw.org/media-centre/news/2020/10/opcw-issues-report-technical-assistance-requested-germany.
5. OPCW, "Investigative science and technology report of the scientific advisory board's temporary working group" (SAB/REP/1/19, OPCW, 2019); https://www.opcw.org/resources/documents/subsidiary-bodies/scientific-advisory-board.
6. R. S. Shaikh, S. J. S. Düsel, B. König, *ACS Catal.* **6**, 8410 (2016).
7. G. G. Rajeshwaran, M. Nandakumar, R. Sureshbabu, A. K. Mohanakrishnan, *Org. Lett.* **13**, 1270 (2011).
8. S. M. A. Kedrowski, D. A. Dougherty, *Org. Lett.* **12**, 3990 (2010).
9. P. Jansa *et al.*, *Green Chem.* **13**, 882 (2011).
10. J. P. Caves Jr., W. S. Carus, "The future of weapons of mass destruction: Their nature and role in 2030" (Occasional Paper 10, National Defense Univ. Press, 2014); https://ndupress.ndu.edu/Portals/97/Documents/Publications/Occasional%20Papers/10_Future%20of%20WMD.pdf.
11. M. G. Beaver *et al.*, *Org. Process Res. Dev.* **24**, 2139 (2020).
12. A. Jasiak *et al.*, *J. Org. Chem.* **84**, 2619 (2019).
13. M. Meanwell, J. Lehmann, M. Eichenberger, R. E. Martin, R. Britton, *Chem. Commun.* **54**, 9985 (2018).
14. IAEA, IAEA safeguards overview: Comprehensive safeguards agreements and additional protocols; https://www.iaea.org/topics/additional-protocol.
15. T. H. Nguyen, *Science* **309**, 1021 (2005).

ACKNOWLEDGMENTS

This article is dedicated to the memory of Ambassador Donald A. Mahley. The author thanks Ambassador R. F. Lehman for constructive feedback. The views and opinions presented here are those of the author and do not necessarily state or reflect those of the US government or Lawrence Livermore National Security, LLC. This work was performed under the auspices of the US Department of Energy by Lawrence Livermore National Laboratory under contract no. DEAC52-07NA27344, document release no. LLNL-JRNL-833178.

10.1126/science.abo6380

TRAVELS FOR AAAS Members and Friends

We invite you to join us!

Discover Mesa Verde & the San Juan Mtns.

September 18-26, 2022

Explore Mesa Verde and dramatic landscapes in SW Colorado and the Canyonlands of SE Utah at the time of golden fall aspen colors.

$3,895 pp + air.

W. AUSTRALIA

Total Solar Eclipse

April 13-24, 2023

We invite you to join our Total Solar Eclipse Adventure to Western Australia, based near Exmouth on 1/4 million acre Bullara Cattle Station. Discover the heritage of Perth and Ningaloo Marine Reserve.

$8,495 pp +air

For a detailed brochure, please call (800) 252-4910

All prices are per person twin share + air

BETCHART EXPEDITIONS Inc.
17050 Montebello Rd
Cupertino, California 95014
Email: AAASInfo@betchartexpeditions.com
www.betchartexpeditions.com

NEWS FROM Science | AAAS

Subscribe to **News from Science** for unlimited access to authoritative, up-to-the-minute news on research and science policy.

bit.ly/NewsFromScience

The tropical ant species *Cyphomyrmex rimosus* farms yeast.

PHOTO: ALEX WILD

AGRICULTURE

Animal agriculturalists

There is much to learn from the farming practices of ants, termites, and other insects

By **Julio C. Postigo**[1] *and* **Joan E. Strassmann**[2]

The chicken, potatoes, and chickpeas on our dinner plates each come from different places where humans transitioned from casual to intense relationships with the animals and plants that sustain us. These transitions happened in at least 12 and probably more than 20 different places 4000 to 21,000 years ago in times of stable and improving climate (*1*).

Some insects have been agriculturalists for tens of millions of years longer than humans. It is therefore possible that we can learn from their agricultural practices to both improve our own and to see what our future might hold. In June 2019, Ted Schultz, Richard Gawne, and Peter Peregrine brought together a team of experts for the 38th Altenberg Workshop in Theoretical Biology to consider this possibility. *The Convergent Evolution of Agriculture in Humans and Insects* synthesizes the outcomes of this event.

The Convergent Evolution of Agriculture in Humans and Insects
Ted R. Schultz, Richard Gawne, Peter N. Peregrine, Eds.
MIT Press, 2022. 338 pp.

The researchers present at the workshop agreed that agriculture is "cultivation on a large scale in which the farmer has become obligately dependent on the cultivated species for nutrition." Agriculture itself they expected to be preceded by cultivation and domestication, where the former simply facilitates growth and proliferation of the crop, be it animal or plant, and the latter requires some form of genetic change that makes the crop more suitable for the farmer and less able to proliferate on its own.

This book discusses many forms of animal agriculture, but its focus is on the relationship between macrotermitine termites and *Termitomyces* fungi more than 24 million years ago and the relationship between attine ants and their fungi more than 55 million years ago. In each of these cases, domestication of the fungus arose only once and caused their hosts to proliferate, making them dominant in their ecosystems.

Perhaps the most interesting message from observing animal rearing of both crops and animals is that the crops are nearly always clonal. This is true for higher termites and leaf-cutting ants, for damselfish that prune their algae gardens, for ambrosia beetles seeding their wood tunnels with fungi, and for ant queens that take their mating flight with a single mealy bug gently held in their mandibles. Clonality eradicates within-crop evolutionary conflict. Such conflict reduces productivity, causing plants to invest more in stems to overtop neighboring plants than in seeds, for example.

The role of clonality or high relatedness among crop plants is something R. Ford Denison explores in depth in chapter 3. While clonality may solve within-crop conflict, it introduces another problem: A single clone will be much more easily defeated by diseases and competitors, although this is less of a problem if individual farmers tend different clones. Other chapters look at the morphological consequences of agriculture for both humans (Menéndez and Buck, chapter 12) and their crops (Gawne and McKenna, chapter 13), emphasizing the complexity of trade-offs.

There are many organisms that depend on symbionts for nutrition. The fungus in a lichen symbiosis depends on the algae or the dinoflagellate it entraps. Coral polyps are animals that depend on plant zooxanthellae for nutrition (*2*). These symbioses between animals or fungi and plants or dinoflagellates involve the cultivation, domestication, and ultimately mutual dependence between two different organisms. These crops are also often clonal for the same reasons as agricultural crops. What, then, constitutes agriculture?

Perhaps the difference between these symbioses and those relationships we call agriculture lies only in that agriculture requires deliberate action and behavior and the other symbioses arise from processes that occur at the physiological or cellular level. But is this really a meaningful difference? If ants and termites cultivated plants rather than fungi, they would not have to perform the behavior of bringing in food, because plants and dinoflagellates grab their carbon from the air.

Termites and attine ants have transformed their ecosystems. But we have taken these transformations even farther, sowing single crops on thousands of acres of the planet. Indeed, there is very little space left for human agricultural expansion (*3*). Perhaps if we paid more attention to how insects farm, we might learn more about reducing competition between domesticated plants and animals, thereby increasing yield and reducing the need for more land. ■

The reviewers are at the [1]Department of Geography, Indiana University Bloomington, Bloomington, IN 47405, USA; and [2]Department of Biology, Washington University in St. Louis, St. Louis, MO 63130, USA. Email: jpostigo@iu.edu; strassmann@wustl.edu

REFERENCES AND NOTES

1. P. H. Kavanagh *et al.*, *Nat. Hum. Behav.* **2**, 478 (2018).
2. A. E. Douglas, *The Symbiotic Habit* (Princeton Univ. Press, 2010).
3. Y. M. Bar-On, R. Phillips, R. Milo, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 6506 (2018).

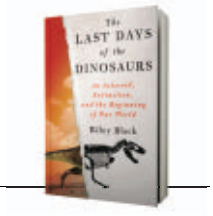
10.1126/science.abq2570

PALEONTOLOGY

Inside the dinosaurs' demise

A science writer offers an accessible account of the Mesozoic's mass extinction

The Last Days of the Dinosaurs: An Asteroid, Extinction, and the Beginning of Our World
Riley Black
St. Martin's Press, 2022. 304 pp.



By **Victoria Arbour**

Sixty-six million years ago, the beginning of the end of the age of dinosaurs came crashing down in the form of a Mount Everest–sized asteroid that struck what is now the Yucatán Peninsula in Mexico. It is a familiar end to many popular dinosaur books, documentaries, and other media published since the 1980s, when the asteroid impact hypothesis began gaining ground as the driver of the end-Cretaceous mass extinction. But the details of this cataclysmic event, and what happened afterward, have rarely received the feature-length treatment that Riley Black presents in *The Last Days of the Dinosaurs: An Asteroid, Extinction, and the Beginning of Our World*.

In just over 300 pages, Black pulls together decades of scientific research on the demise of the nonavian dinosaurs into a deeply compelling narrative of both luck and misfortune in the face of almost unimaginable calamity. Major advances over the past 15 years in particular have enhanced our understanding of the end of the Mesozoic era and its mass extinction, making this a timely and distinctive addition to the ever-expanding landscape of popular science dinosaur books.

Starting just a few days before the asteroid impact, Black centers the story on the animals present in the Hell Creek Formation of the western United States, a geological unit that has been extensively studied for more than 100 years and that provides the best window we have into the time before and after the mass extinction. Through the eyes of the inhabitants of Hell Creek, readers pass through the moment of impact and then the first hour, day, month, and year of the Cenozoic era.

The story continues with chapters set one hundred, one thousand, one hundred thousand, and one million years after the asteroid impact, charting the slow but steady reshaping of the world. The final chapter includes a visit to a geological site preserving markers of the impact itself and reflections on the nature of mass extinction and what lessons humanity can take away from this singular event. Each chapter ends with a detour to somewhere else on the planet—Antarctica, India, and the Atlantic Ocean, for example—providing a global balance to this otherwise tightly focused narrative.

A marsupial traverses a branch above a *Pachycephalosaurus* in this artistic rendering of the dinosaurs' last days.

Unlike the other mass extinctions recorded in the fossil record, extinction for most species at the end of the Cretaceous probably happened within a few hours or days after the asteroid impact. Black's writing captures the horror of the impact event without lingering on doom and gloom, instead keeping a steady momentum focusing on survival and change. *The Last Days of the Dinosaurs* emphasizes the limits of adaptation and natural selection in the face of abrupt environmental change, highlighting how this singular event created a powerful filter that favored the small and unfussy over large and specialized creatures.

Many of the main players in the book's first few chapters—*Tyrannosaurus*, *Triceratops*, and *Ankylosaurus*, to name a few—will be familiar to most readers. But the animals often overshadowed (metaphorically and literally) by dinosaurs are given some much-deserved attention here, too. Black recounts stories of lucky survivors whose descendants will be familiar to many readers—frogs, snakes, turtles, and crocodiles—and organisms that rarely get a mention in dinosaur stories, such as the planktonic coccoliths and coil-shelled ammonites. And of course, as the nonavian dinosaurs pass into history, we see the changes that happen to the ancestors of today's mammals, who evolved into new forms in the empty spaces created by the asteroid's aftermath.

Instead of extensively annotating the text with footnotes or citations, Black provides a welcome appendix summarizing the evidence used to build the stories told in each chapter. This keeps the text flowing without interruptions while still making clear the limits of current scientific evidence and where informed speculation fills in the gaps. Detailed notes close out the book, with references to primary literature for those who want to dig deeper. Each chapter also includes charming illustrations of that chapter's main "protagonist" by artist Kory Bing.

The Last Days of the Dinosaurs would fit equally comfortably on the bookshelf of a die-hard dinosaur enthusiast, someone revisiting their childhood love of dinosaurs or paleontology, and anyone interested in the science of extinction or the transition from the age of dinosaurs to the age of mammals. ■

The reviewer is curator of paleontology at the Royal BC Museum, Victoria, BC V8W 9W2, Canada. Email: victoria.arbour@gmail.com

10.1126/science.abo7409

IMAGE: JULIUS T. CSOTONYI/SCIENCE SOURCE

Floating wind turbines, such as these two en route to the world's first floating wind farm, could affect the environment in ways that have not yet been identified.

PHOTO: TERJE AASE/SHUTTERSTOCK

Edited by **Jennifer Sills**

Offshore renewables need an experimental mindset

The development of floating wind turbines that can operate in deep, offshore waters has unlocked tremendous energy generation potential (*1*). Existing floating offshore wind turbines, however, are still in demonstration phases. Because only about 10 turbines exist worldwide (*2*), their short- and long-term environmental impacts are still largely unknown. Floating wind turbines are likely to come with their own set of unique risks (*3*), which could include secondary entanglement of marine life in debris ensnared on stabilizing mooring lines (*4*), increased collision potential due to three-dimensional turbine movement (*5*), and benthic habitat degradation from turbine infrastructure such as anchors and buried interarray cables (*6*).

Despite potential impacts, countries are rapidly moving toward full commercial installations. The United States is advancing toward a lease sale for two areas in central and northern California and proposing floating wind turbines as a primary technology for the Gulf of Mexico (*7*). Floating wind turbines are also planned for the Gulf of Maine (*8*) and likely for New York (*9*). European and Asian countries have similar expansions planned (*2*).

Countries need robust plans to prevent, monitor, and mitigate the environmental impacts of floating wind turbines. We urge energy authorities and lawmakers to treat each installation as an experiment to gather information about the costs and benefits of this fledgling technology (*10*). Like any experiment, a comprehensive monitoring scheme is required to collect data, ideally for several years before turbines are first placed and then through the construction, lifetime operations, and decommissioning of the turbines (*11*). A robust monitoring plan with funding secured across all phases will help distinguish effects of floating wind development from other factors, such as climate change. Although it is tempting to focus only on the positives of clean energy, it is crucial to think preemptively about the longer-term impacts of floating wind turbines and use adaptive management practices to minimize impacts accordingly if necessary (*12*). Prevention rather than cure will be essential for the long-term sustainable success of this exciting, yet unknown, new sector.

Andrew F. Johnson[1,2]*, Cyndi L. Dawson[3], Melinda G. Conners[4], Cameron C. Locke[5], Sara M. Maxwell[5]

[1]MarFishEco Fisheries Consultants Ltd, Edinburgh, Scotland, UK. [2]Marine Sustainability, Policy & Conservation Evidence (Marine SPACE) Group, The Lyell Centre, Institute of Life and Earth Sciences, School of Energy, Geoscience, Infrastructure and Society, Heriot-Watt University, Edinburgh, Scotland, UK. [3]Castalia Environmental, Santa Cruz, CA 95062, USA. [4]School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, NY 11794, USA. [5]School of Interdisciplinary Arts and Sciences, University of Washington, Bothell, WA 98011, USA.
*Corresponding author.
Email: andrew@marfisheco.com

REFERENCES AND NOTES

1. P. Rosa-Aquino, "Floating wind turbines could open up vast ocean tracts for renewable power," *The Guardian* (2021).
2. J. Lee., F. Zhao, "Global Offshore Wind Report," *Global Wind Energy Council* (2021).
3. S. M. Maxwell *et al*., *J. Environ. Manage*. **307**, 114577 (2022).
4. S. Benjamins *et al.*, "Understanding the potential for marine megafauna entanglement risk from marine renewable energy developments," Scottish Natural Heritage Commissioned Report No. 791 (2014), p. 95.
5. H. Bailey, K. L. Brookes, P. M. Thompson, *Aquat. Biosyst*. **10**, 8 (2014).
6. H. K. Farr *et al*., *Ocean Coast. Manage*. **207**, 105611 (2021).
7. Bureau of Ocean Energy Management, "BOEM hosts second Gulf of Mexico Renewable Energy Task Force meeting" (2022); www.boem.gov/newsroom/notes-stakeholders/boem-hosts-second-gulf-mexico-renewable-energy-task-force-meeting.
8. State of Maine Governor's Energy Office, "Gulf of Maine floating offshore wind research array" (2021); www.maine.gov/energy/initiatives/offshorewind/researcharray.
9. New York State Energy Research and Development Authority, "Governor Hochul announces nation leading $500 million investment in offshore wind" (2022); www.nyserda.ny.gov/About/Newsroom/2022-Announcements/2022-01-05-Governor-Hochul-Announces-Nation-Leading-500-Million-Investment-in-Offshore-Wind.
10. B. Snyder, M. J. Kaiser, *Renew. Energ*. **34**, 1567 (2009).
11. A. Giron-Nava *et al*., *Mar. Ecol. Prog. Ser*. **572**, 269 (2017).
12. A. Copping, V. Gartman, R. May, F. Bennet, in *Wind Energy and Wildlife Impacts: Balancing Energy Sustainability with Wildlife Conservation*, R. Bispo, J. Bernardino, H. Coelho, J. Lino Costa, Eds. (Springer International Publishing, 2019), pp. 1–25.

COMPETING INTERESTS

A.F.J. was funded by the Natural Resources Defense Council to consult on the environmental and fishery impacts of floating offshore wine turbines.

10.1126/science.abo7924

Green energy threatens Chile's Magallanes Region

On 2 December 2021, Chile's minister of energy and mining announced the country's largest green hydrogen project, to be developed in Chile's southernmost Magallanes Region (*1*–*3*). The project is intended to help achieve Chile's stated goal of generating 25 GW of green hydrogen by 2030 (*1*, *4*). However, enthusiasm for clean energy projects obscures their environmental and cultural impacts.

Despite the potential benefits, the large scale of this green hydrogen megaproject,

particularly its wind farms, could have an outsized effect on both ecological processes and the surrounding landscape. San Gregorio and Tierra del Fuego form part of important migration routes of threatened birds such as the ruddy-headed goose, the red knot, and the Magellanic plover (*5*), which fly across Patagonia on their way to their austral summer areas. Replacing sheep ranching with wind generation plants also entails a profound cultural change, comparable to the changes brought about by the arrival of European immigrants and inhabitants of the Chiloé archipelago and the subsequent development of sheep ranching in Magallanes at the end of the 19th century, which reconfigured social relations and land use in the region (*6*, *7*).

Preliminary estimates based on a pilot project in Punta Arenas (*3*) suggest that the megaproject could reach about 2900 installed wind turbines by 2027, occupying an area of at least 150,000 hectares. This would represent a 320% increase in Chile's wind energy generation capacity and would represent 1.35% of the wind energy installed in the world [relative to 2021 data (*8*)]. Recent studies in central Chile show a rate of 0.6 to 1.8 bird collisions per wind turbine per year (*3*). Scaling this to the magnitude of the planned Magallanes project could lead to between 1740 and 5220 bird collisions per year. However, this estimate does not consider that the Magallanes Region is a migration area for about 43 species of birds, including Passeriformes, Charadriiformes, and Strigiformes (*5*, *9*), which would likely increase these numbers.

Environmental impact assessments of these projects must take into consideration the high natural value of this landscape, with protected areas such as Torres del Paine National Park, Pali Aike National Park, and Bahía Lomas Ramsar site and Nature Sanctuary (*10*). Failing to do so could turn the development of clean energy megaprojects into another example of extractivist development (*11*), which would export a product (green hydrogen) to Europe and Asia while generating potentially irreversible changes to the local environment and culture.

Heraldo V. Norambuena[1]*, Fabio A. Labra[2], Ricardo Matus[1,3], Humberto Gómez[4], Diego Luna-Quevedo[5], Carmen Espoz[1]

[1]Centro Bahía Lomas, Facultad de Ciencias, Universidad Santo Tomás, Concepción Chile. [2]Centro de Investigación e Innovación para el Cambio Climático, Facultad de Ciencias, Universidad Santo Tomás, Santiago, Chile. [3]Centro de Rehabilitación de Aves Leñadura, Punta Arenas, Chile. [4]Agrupación Ecológica Patagónica, Punta Arenas, Chile. [5]Western Hemisphere Shorebird Reserve Network Executive Office–Manomet, Plymouth, MA 02360, USA.
*Corresponding author.
Email: hnorambuena@santotomas.cl

REFERENCES AND NOTES

1. Ministerio de Energía, "El más grande de Chile: Ministro Jobet anuncia nuevo proyecto de hidrógeno verde en Magallanes" (2021); https://energia.gob.cl/noticias/nacional/el-mas-grande-de-chile-ministro-jobet-anuncia-nuevo-proyecto-de-hidrogeno-verde-en-magallanes [in Spanish].
2. Highly Innovative Fuels, "Capítulo 1: Descripción de proyecto—Proyecto piloto de descarbonización y producción de combustibles carbono neutral, declaración de impacto ambiental," Tech. Rep. N° 2020-12-31 (2020); https://infofirma.sea.gob.cl/DocumentosSEA/MostrarDocumento?docId=6e/7c/7dcdd2b1c3afdfbd2f257113b0281c634710 [in Spanish].
3. República de Chile, Comisión de Evaluación, Región de Magallanes y Antártica Chilena, "Califica ambientalmente el proyecto: Proyecto piloto de descarbonización y producción de combustibles carbono neutral" (Resolución de Calificación Ambiental N°58, 2021) [in Spanish].
4. Ministerio de Energía, "Transición energética de Chile, política energética nacional" (Gobierno de Chile, 2021) [in Spanish].
5. F. Medrano, R. Barros, H. V. Norambuena, R. Matus, F. Schmitt, Eds., *Atlas de las Aves Nidificantes de Chile* (Red de Observadores de Aves y Vida Silvestre de Chile, Santiago, Chile, 2018) [in Spanish].
6. J. Calderón, *Boletín del Ministerio de Agricultura* **10**, 1 (1936) [in Spanish].
7. R. Urbina, in *Planning Outlook Series 1* (1956), vol. 4, p. 22.
8. Global Wind Energy Council, Global wind report 2021 (2021); https://gwec.net/global-wind-report-2021/.
9. eBird: An online database of bird distribution and abundance [web application] (Cornell Lab of Ornithology, Ithaca, New York, 2022); https://ebird.org/chile/barcharт?byr=1900&eyr=2022&bmo=1&emo=12&r=CL-MA.
10. Comisión Nacional del Medio Ambiente Chile, "Diagnóstico y propuesta para la conservación de la biodiversidad en la xii región" (2002) [in Spanish].
11. M. A. Urbina *et al.*, *Science* **373**, 1208 (2021).

10.1126/science.abo4129

Brazilian pesticides law could poison the world

Brazil's National Congress will soon vote on a controversial bill (PL 6299/2002) that relaxes the current legislation on pesticides (*1*). Arguing that the registration of new products takes too long, this bill proposes changes to the evaluation and authorization process, excluding the health and environment federal agencies from the decision. In addition, previously banned substances could then be reevaluated under these new rules. This bill fits Brazil's recent trend of undermining environmental law (*2*) by prioritizing the productive sector to the detriment of environmental integrity (*3*, *4*).

In 2021, the government authorized the use of 562 new agrochemicals in Brazil (*5*), many of them imported from Europe and North America (*6*). Several of those new pesticides are banned in these countries (*6*, *7*), but their manufacturers continue exporting them to places with permissive legislation like Brazil. The indiscriminate use of pesticides without proper evaluation is a matter of public health. In the past 10 years, intoxication and deaths related to pesticide poisoning increased by 94% in Brazil (*8*), and those pesticides persist in the environment (*9*).

Because Brazil is a leader in exporting its crops, such as soy that supplies global animal feed (*10*), the likely approval of this bill should be a global concern. More pesticides are not necessary to feed the world (*11*). There are well-known solutions to enhance productivity (*12*) that do not require the intense use of pesticides, such as agroecology (*11*). An alternative bill (PL 6670/2016) could move Brazil in a better direction by initiating a national program to reduce pesticides, but this proposal has been given low priority and is unlikely to become law under the current administration. Strengthening environmental agencies and investing in science and technology is the way to achieve the sustainable development of agribusiness.

Laís Carneiro[1]*, Larissa Faria[1], Natali Miiller[1], André Cavalcante[1], Afonso Murata[2], Jean Ricardo Simões Vitule[1]

[1]Laboratório de Ecologia e Conservação, Setor de Tecnologia, Departamento de Engenharia Ambiental, Universidade Federal do Paraná, Curitiba, PR, 81531-970, Brazil. [2]Centro de Ensino Pesquisa e Extensão em Agroecologia, Departamento de Fitotecnia e Fitossanidade, Universidade Federal do Paraná, Curitiba, PR, 81531-970, Brazil.
*Corresponding author.
Email: lais.olicar@gmail.com

REFERENCES AND NOTES

1. Agência de Notícias, "Câmara aprova projeto que altera regras de registro de agrotóxicos." (2022); www.camara.leg.br/noticias/849479-camara-aprova-projeto-que-altera-regras-de-registro-de-agrotoxicos/ [in Portuguese].
2. L. G. Barbosa, M. A. S. Alves, C. E. V. Grelle, *Land Use Pol.* **104**, 105384 (2021).
3. P. Charvet *et al.*, *Science* **371**, 356 (2021).
4. F. E. Coelho, L. C. Lopes, R. M. S. Cavalcante, G. C. Corrêa, A. O. H. C. Leduc, *Science* **365**, 552 (2019).
5. Brazil, Ministry of Agriculture, "Registros concedidos 2000–2022" (2022); www.gov.br/agricultura/pt-br/assuntos/insumos-agropecuarios/insumos-agricolas/agrotoxicos/RegistrosConcedidos20002022.xlsx [in Portuguese].
6. L. M. Bombardi, *Geografia do Uso de Agrotóxicos No Brasil e Conexões com a União Europeia* (Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, 2017) [in Portuguese].
7. "PAN International consolidated list of banned pesticides," Pesticide Action Network International (2021); https://pan-international.org/pan-international-consolidated-list-of-banned-pesticides/.
8. R. J. Buralli, F. N. E. F. de Souza, "Mortality and morbidity by work-related pesticide poisoning in Brazil, 2009–2019," ISEE Conference Abstracts 2021 (2021).
9. I. E. Barnhoorn, M. S. Bornman, C. J. Van Rensburg, H. Bouwman, *Chemosphere* **77**, 1236 (2009).
10. C. Dowler, "Soya, corn, and cotton make Brazil world leader for hazardous pesticides," *Unearthed* (2020); https://unearthed.greenpeace.org/2020/02/20/brazil-pesticides-soya-corn-cotton-hazardous-croplife/.
11. "Report of the Special Rapporteur on the right to food," United Nations A/HRC/34/48 (2017).
12. M. Lykogianni, E. Bempelou, F. Karamaouna, K. A. Aliferis, *Sci. Tot. Environ.* **795**, 148625 (2021).

10.1126/science.abo6942

Pushing the Boundaries of Knowledge

As AAAS's first multidisciplinary, open access journal, *Science Advances* publishes research that reflects the selectivity of high impact, innovative research you expect from the *Science* family of journals, published in an open access format to serve a vast and growing global audience. Check out the latest findings or learn how to submit your research: **science.org/journal/sciadv**

Science Advances
AAAS

GOLD OPEN ACCESS, DIGITAL, AND FREE TO ALL READERS

RESEARCH

IN *SCIENCE* JOURNALS

Edited by **Michael Funk**

CELL BIOLOGY

Cell division facilitates cell invasion

Cell migration through compact tissues within multicellular organisms is essential for normal embryonic development, immune response, and cancer metastasis. Frequently, migrating cells find a path of least resistance; however, relatively little is known about invasion into cell-dense tissues with no preexisting paths. Studying fruit flies, Akhmanova *et al.* discovered that embryonic macrophages can penetrate into tissues when at least one surrounding tissue cell divides at the site of invasion. Live imaging and genetic and optogenetic manipulations revealed that macrophage invasion requires the release of intercellular adhesions, which occurs during division. —BAP *Science*, abj0425, this issue p. 394

Artist's depiction of a macrophage penetrating a tissue layer at a site of cell division

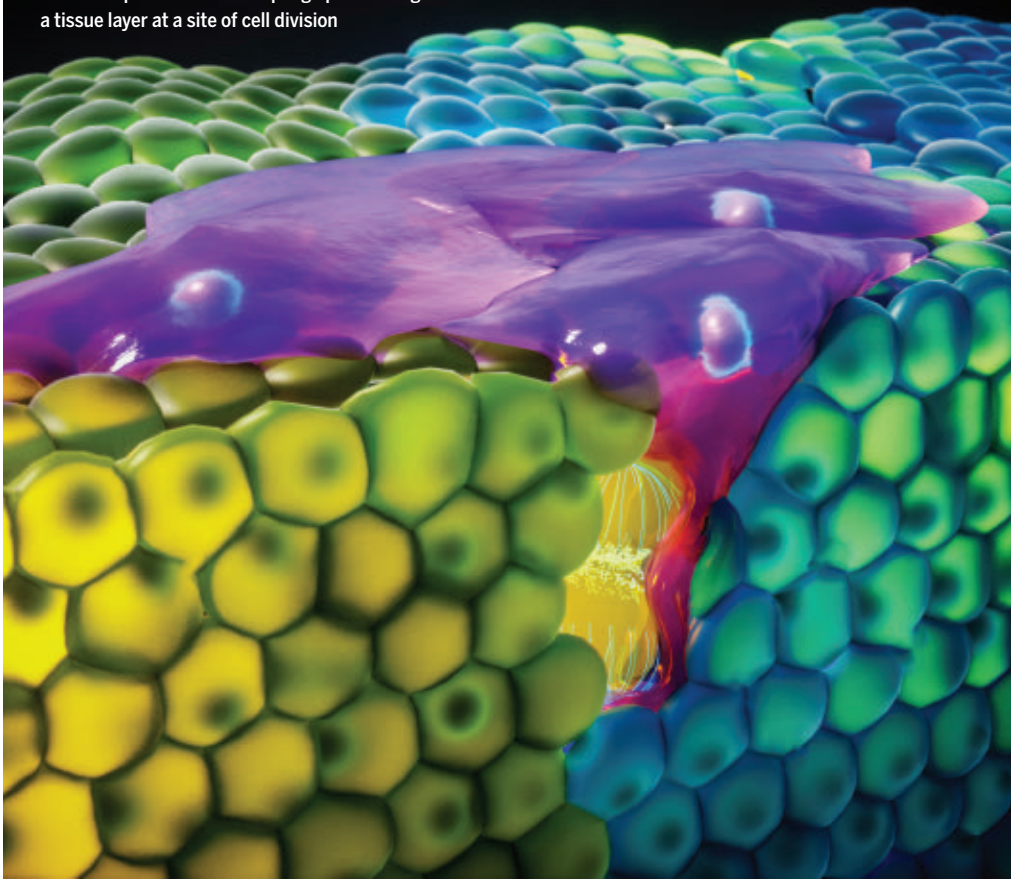


IMAGE: ELLA MARU STUDIOS, INC.

ELECTROOPTICS

Breaking down the domain walls

Ferroelectric materials should make for decent optical components because of the change in refractive index with electric field. However, many of the best ferroelectrics have domain walls that scatter light and are not useful for optics applications. Liu *et al*. used a high-temperature poling method to remove the light-scattering domain walls in a lead ceramic ferroelectric. The material has a very high electro-optic coefficient and requires a very low driving voltage. This strategy may be useful for other materials and may help in the development of better optical devices. —BG *Science*, abn7711, this issue p. 371

PROTEIN DESIGN

Steps toward a nanomachine

Protein rotary machines such as ATP synthase contain axle-like and ring-like components and couple biochemical energy to the mechanical work of rotating the components relative to each other. Courbet *et al*. have taken a step toward designing such axle-rotor nanomachines. A structural requirement is that interactions between the components must be strong enough to allow assembly but still allow different rotational states to be populated. The authors met this design challenge and computationally designed ring-like protein topologies (rotors) with a range of inner diameters that accommodate designed axle-like binding partners. The systems assemble and populate the different rotational states anticipated by the designs. These rotational energy landscapes provide one of two needed elements for a directional motor. —VV *Science*, abm1183, this issue p. 383

ORGANIC CHEMISTRY

Better C–C coupling through ligand swaps

The development of cross-coupling catalysis for carbon–carbon bond formation revolutionized pharmaceutical synthesis. Nonetheless, one drawback of the original reactions is the need to activate one of the coupling partners ahead of time. Recently, chemists have focused on direct coupling of two halocarbons, which is efficient but poses a selectivity challenge. Hamby *et al*. report a nickel catalyst for alkyl-aryl coupling that relies on ligand exchanges in concert with electrochemistry to react with each partner consecutively and thereby avoid alkyl-alkyl, aryl-aryl, or isomerized by-products. —JSY *Science*, abo0039, this issue p. 410

IN OTHER JOURNALS

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

QUANTUM SENSING

Spatial sensing with hydrogen

The coupling of coherent quantum systems to the surrounding environment can be used to create highly sensitive sensors, but many surface-based implementations, such as nitrogen vacancies in diamond, have limited spatial resolution. Wang *et al*. created a quantum sensor with high spatial sensitivity by using the electric field from a scanning tunneling microscope tip to confine a hydrogen molecule on copper nitride islands grown on a copper surface. Femtosecond-pulse terahertz spectroscopy was used to follow the coherence of the two-level system created by different adsorption geometries. The temporal oscillations and decoherence in the superposition state showed spatial variations at the sub-angstrom scale. —PDS

Science, abn9220, this issue p. 401

BIOMATERIALS

Designing degradable dressings

Synthetic scaffolds are an attractive option for dressings to treat skin wounds because they are shelf-stable, have tunable and defined chemical compositions, and are more affordable than naturally derived scaffolds or cell-based dermal substitutes. Patil *et al.* investigated the role of scaffold hydrophilicity in polythioketal-based polyurethane (PTK-UR) foam scaffold resorption and promotion of tissue regeneration in wound healing. They discovered that hydrophilic scaffolds with seven ethylene glycol units between the thioketal bonds in the polymer backbone exhibited optimal reactive oxygen species–dependent degradation and porcine skin wound healing, including ischemic flap excisional wounds. These results support further investigation of PTK-UR formulations as alternatives to commercially available wound matrices. —CC

Sci. Transl. Med. **14**, eabm6586 (2022).

OPTOELECTRONICS

Probing interlayer excitons

Stacking of atomically thin layers of van der Waal materials gives rise to a wealth of exotic transport properties. Controlling the twist angle and stacking sequence can yield heterostructures exhibiting magnetism, superconductivity, and the generation of interlayer excitons (electron-hole pairs across the layers). Barré *et al*. developed an electromodulation spectroscopic technique with which the detailed band structure can be determined, providing a key to understanding how the exotic optoelectronic transport properties develop in these heterostructure materials. —ISO

Science, abm8511, this issue p. 406

IMAGING TECHNOLOGY

Portable MRI for ischemic stroke

Magnetic resonance imaging (MRI) offers unparalleled soft tissue contrast for diagnostic imaging, but the instrumentation is bulky and cannot be transported. Portable MRI systems could be lower cost and useful in trauma settings or at the bedside, especially in situations where time is of the essence. Portability, however, comes at the cost of lower magnetic field strengths, motion artifacts, and associated performance loss. Successful deployment of point-of-care MRI in such environments represents a formidable engineering challenge. Yuen *et al.* demonstrate the feasibility of this technology with sufficient image quality and contrast for brain imaging (see the Focus by Basser), providing evidence for the usefulness of this technology in the context of assessing ischemic stroke. —LSB

Sci. Adv. 10.1126/sciadv.abm3952, 10.1126/sciadv.abp9307 (2022).

PHOTO: MAURUSONE/ISTOCK.COM

MATERIALS SCIENCE

Scorpion-inspired strain sensors

Most strain sensors are designed to be highly sensitive to movement in one direction by using long uniaxial sensor elements while being insensitive to motion in other directions. To compensate for their poor vision, scorpions have developed powerful vibration-sensing organs called sensillum on their walking legs. Liu *et al.* drew on these as inspiration for designing flexible, highly sensitive omnidirectional strain sensors. A film with cut, curved grooves was placed on a poly(dimethylsiloxane) layer and coated with silver nanoparticles so that the grooves formed around a central ring. This ensures that the stress concentration is not at the focus of the grooves and that some of the segments will deform maximally to generate omnidirectional strain sensing. —MSL

Adv. Mat. 10.1002/adma.202200823 (2022).

Scorpions, like the forest scorpion *(Heterometrus petersii)* pictured, have flexible strain-sensors that have inspired human-engineered versions.

STRUCTURAL BIOLOGY

The molecular effects of oxytocin

Oxytocin is a signaling peptide that binds to a G protein–coupled receptor to regulate diverse physiological processes in humans. Oxytocin is perhaps best known for its roles in causing uterine contractions during labor, initiating nurturing maternal behavior, and affecting the mother's ability to nurse. Meyerowitz *et al.* report a structural analysis of the native state of the oxytocin receptor bound to oxytocin and a model G protein. The structure, together with functional assays, shows that binding of oxytocin disrupts the TM7 helix in the receptor, and this effect is required for full activation. The authors also identified a cation-coordinating residue that makes activation of the oxytocin receptor dependent on magnesium. —VV

Nat. Struct. Mol. Biol. **29**, 274 (2022).

CELESTIAL MECHANICS

How did Ceres get there?

The dwarf planet Ceres is the largest object in the asteroid main belt. It has an ice-rich composition, which suggests that it originally formed in the outer Solar System. Ribeiro de Sousa *et al.* performed orbital

MIGRATION

Father knows best

Many birds migrate long distances between breeding and wintering grounds. Although we know that in many cases naïve birds follow more experienced birds, we still know little about the details of which one is doing the leading. Byholm *et al.* monitored migration in Caspian terns using GPS trackers and found that family groups generally migrated together, with a single parent leading the way. Both male and female parents led the group, but fathers were more likely to be guides. Individuals that lost their guides tended to fall victim to birds of prey. The authors did report that some fathers would act as guides for unrelated juveniles. When migrating solo for the first time, subadult terns followed the same route they took with their parents, indicating cultural inheritance of migration knowledge. —SNV *Nat. Commun.* **13**, 1566 (2022).

Caspian tern (*Hydroprogne caspia*) juveniles rely on parental guidance, chiefly from their fathers, to ensure safe migration.

dynamics simulations to test this scenario. They found that migration of the giant planets in the early Solar System could scatter numerous planetesimals into the main belt. Most were then ejected by interactions with Jupiter, with only a small fraction evolving onto stable orbits. To explain one Ceres-sized object surviving in the main belt, there must have originally been several thousand in the outer Solar System. This is consistent with estimates based on the formation of Uranus and Neptune. —KTS

Icarus **379**, 114933 (2022).

ORGANIC CHEMISTRY

Ladders in the mirror

Ammonia-oxidizing bacteria produce a fascinating class of lipids that have multiple edge-sharing four-carbon rings in their tails. These molecules have been called ladderanes because when they are drawn flat, the shared edges look like the rungs of a ladder. However, in three-dimensions, the geometry is more reminiscent of a stairway. Ray *et al.* report a synthesis of ladderanol, an alcohol-capped component of these lipids that selects between the two product mirror images (natural and unnatural) at a late stage. They achieved the requisite selectivity using a hydrogen-bonding squaramide catalyst to guide a nitroalkyl Michael addition. —JSY

Angew. Chem. Int. Ed. 10.1002/anie.202201584 (2022).

IMMUNOMETABOLISM

Steely response to cancer

Ferroptosis is a form of programmed cell death that requires iron and the accumulation of lipid peroxides. Cancer cells undergo ferroptosis in response to immunotherapy, but how this process is regulated is not clear. Liao *et al.* explored whether factors within the tumor microenvironment were involved in tumor ferroptosis. They studied specific lipids and lipid metabolic enzymes that influence antitumor immunity and found that arachidonic acid synergized with T lymphocyte–derived interferon-γ to produce pro-ferroptotic effects. A combination of PD-L1 immunotherapy and arachidonic acid supplementation was tested as a preclinical treatment strategy to sensitize tumors toward ferroptosis and reduce tumor growth in mice. —PNK

Cancer Cell **40**, 365 (2022).

NEURODEVELOPMENT

Tuning layers of the brain

The cortex of the adult mammalian brain is made up of six layers of neurons, but not all regions of the brain show layers of similar depth and complexity. Layer 4, which receives sensory information from the thalamus, is thicker in the sensory cortex than it is in the motor cortex of the brain. Sato *et al.* examined communication between thalamocortical axons and the developing cortex in mice. Postnatal ablation of thalamocortical axons showed how dependent the normal development of layer 4 is on these neuronal connections. Without thalamocortical axon connections, layer 4 in the somatosensory cortex had fewer neurons than it normally would. Even minute alterations in the laminar architecture of the brain's cortex can have consequences for cognitive, mental, or psychiatric disorders. —PJH

eLife **11**, e67549 (2022).

PHOTO: MICHAEL WOODRUFF/SHUTTERSTOCK

ALSO IN *SCIENCE* JOURNALS

Edited by **Michael Funk**

GENETICS

Can repetitive DNA cause disease?

Repetitive DNA sequences litter the human genome. Although they are mostly silent, some can "jump" around, duplicating and inserting themselves throughout genomic DNA, which can potentially disrupt nearby genes, modulate transcription, or elicit immune responses. This insertional mutagenesis is especially prominent in cancer and has also been described in developmental, degenerative, and autoimmune diseases. However, it is unclear whether this process contributes to pathogenesis or if it arises as a consequence of other genetic and epigenetic changes. In a Perspective, Burns considers the evidence and discusses mechanisms for insertional mutagenesis in pathogenesis and whether repetitive DNA is a widespread cause of disease. —GKA

Science, abl7399, this issue p. 353

TRAFFIC SAFETY

When behavioral nudges fail

Do traffic safety interventions work? Hall and Madsen present evidence from a study in Texas showing that the number of crashes actually increases by a few percentage points when motorists are confronted with displays indicating the number of road fatalities in the area (see the Perspective by Ullman and Chrysler). The authors suggest that this counterintuitive finding results from a cognitive overload experienced by drivers when confronted with multiple notices and instructions on complex stretches of road, leading to distraction. They conclude that traffic safety "nudges" need to be carefully designed and positioned to avoid backfiring. —AMS

Science, abm3427, this issue p. 370; see also abq1757, p. 347

CELL BIOLOGY

Membrane repair in target cell defenses

Killer T cells destroy virus-infected and cancer cells by secreting two protein toxins that act as a powerful one-two punch. Pore-forming toxins, perforins, form holes in the plasma membrane of the target cell. Cytotoxic proteins released by T cells then pass through these portals, inducing target cell death. Ritter *et al*. combined high-resolution imaging data with functional analysis to demonstrate that tumor-derived cells fight back (see the Perspective by Andrews). Protein complexes of the ESCRT family were able to repair perforin holes in target cells, thereby delaying or preventing T cell–induced killing. ESCRT-mediated membrane repair may thus provide a mechanism of resistance to immune attack. —SMH

Science, abl3855, this issue p. 377; see also abp8641, p. 346

CATALYSIS

Sourcing hydrogen with ceramic reactors

An alternative to directly transporting hydrogen produced at large scales through steam reforming for applications such as vehicular fueling is smaller-scale, on-site production from methane or carriers such as ammonia. The hydrogen produced must be separated from co-produced carbon dioxide or nitrogen. Proton ceramic electrochemical reactors can extract pure hydrogen from gas mixtures by electrolytically pumping protons across the membrane at 800°C, but as the extraction proceeds, temperature gradients and entropic effects lead to efficiency drops. Clark *et al*. developed a nickel-based glass-ceramic composite interconnect that allowed for the design of a more complex reactor pathway (see the Perspective by Shih and Haile). Counterflowing streams balanced heat flows and maintained stable operating conditions that enabled 99% efficiency of hydrogen recovery. —PDS

Science, abj3951, this issue p. 390; see also abo5369, p. 348

SUPERCONDUCTIVITY

Double horn points to an exotic state

The so-called Fulde-Ferrell-Larkin-Ovchinnikov (FFLO) state, a particular kind of superconducting order in which the order parameter oscillates in space, has fascinated physicists because it is tricky to stabilize. Kinjo *et al*. used nuclear magnetic resonance (NMR) measurements to observe a signature of the FFLO state in the material strontium ruthenate (see the Perspective by Pavarini). The modulation in the superconducting order parameter caused a corresponding modulation of spin density, which resulted in a peculiar, double horn–shaped structure of NMR intensity for a certain range of temperatures and applied magnetic field. —JS

Science, abb0332, this issue p. 397; see also abn3794, p. 350

METAMATERIALS

Actively structuring light

The development of metasurfaces has provided a route to replacing bulk optical components with thin layers of engineered materials. In a Review, Dorrah and Capasso highlight some of the recent advances in wavefront shaping using multifunctional metaoptics. They focus on the ability to tune the response of the metasurface by simply tuning one or more degrees of freedom of incident light, for example, by varying its angle of incidence, polarization, wavelength, or phase. The key feature of these metasurfaces is that although they are static, they can produce a tunable response without the need for complex switching. These developments enable multifunctional and lightweight components for technologies such as augmented and virtual reality displays, drone-based sensing, and endoscopy. —ISO

Science, abi6860, this issue p. 367

CANCER GENOMICS

A signature accomplishment

Tumor development is associated with the accumulation of mutations in the genome. Depending on the causes of a given cancer, such as environmental exposures or DNA repair abnormalities, these mutations can form a specific pattern called a mutational signature. Many mutational signatures have already been reported in cancer, but by performing whole-genome sequencing on a particularly large collection of cancer samples, Degasperi *et al*. not only confirmed previously reported signatures, but also discovered many rarer ones (see the Perspective by Szüts). The authors characterized these signatures, tried to elucidate the underlying biology where possible, and then provided an algorithm for applying these findings to individual patients to help personalize cancer treatments. —YN

Science, abl9283, this issue p. 368; see also abo7425, p. 351

CORONAVIRUS

Acting against Omicron

The rapid spread of the Omicron variant of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus was partly due to its ability to evade antibodies elicited against the viral spike protein of early variants. Zhou *et al*. identified antibodies that maintained their ability to neutralize the Omicron

variant and used functional and structural studies to determine how these antibodies recognize the Omicron spike protein. Their studies identify antibody combinations with therapeutic potential and could inform the development of vaccines. —VV

Science, abm8897, this issue p. 369

CANCER THERAPY

Sensitizing AML to kinase inhibitors

Although some acute myeloid leukemia (AML) cells have mutations that activate the MEK kinase pathway, MEK inhibitors are generally ineffective in treating this cancer. Using AML cells from patients and cell lines, Pedicona *et al*. correlated kinase network activity and epigenetic markers (see the Focus by Narayan and Huntly). Inhibiting the epigenetic modifying enzyme LSD1 rewired kinase activity in AML cells and sensitized them to subsequent MEK inhibitor treatment. Thus, activated MEK signaling might be a source of resistance to LSD1 inhibitors currently being tested in AML patients. —LKF

Sci. Signal. **15**, eabl7989, eabo0059 (2022).

FISHERIES

Pressure from both sides

It is now well established that human activities impose selection that leads to evolutionary change in wild species. Some of the best examples of this process are from harvested species, especially fishes, and we know that large-scale fishing has led to effects such as early maturation and changes in adult size. Czorlich *et al*. looked at native populations of Atlantic salmon and found that even indigenous fishing pressure has led to evolutionary change in these populations (see the Perspective by Therkildsen and Pinsky). They also identified an additional indirect source of selection: harvest of salmon prey species at sea. —SNV

Science, abg5980, this issue p. 420; see also abo6512, p. 344

SKIN INFLAMMATION

Gene interface tool to identify your rash

Many chronic skin conditions are caused by imbalances in the immune milieu of the environment, yet many of these changes remain undercharacterized. Liu *et al.* used single-cell profiling of immune cells from skin samples of atopic dermatitis, psoriasis vulgaris, and healthy controls to profile the immune signatures of the different diseases. They found that atopic dermatitis and psoriasis vulgaris samples had distinct resident memory T cell, innate lymphoid cell, and $CD8^+$ T cell gene signatures that could be validated in an external dataset. Using these signatures, they created a tool that could classify previously clinicopathologically indeterminate rash samples. This tool is now available as a web resource (RashX) for anyone to use to help further classify skin rashes. —DAE

Sci. Immunol. **7**, eabl9165 (2022).

MICROBIOLOGY

Fungal pathogen circumvents cellulose

Plant cell walls impede colonization by fungal pathogens, and crystalline cellulose has long been regarded as the strongest component of these walls. Studying the plant pathogen *Fusarium oxysporum*, Gámez-Arjona *et al.* found that deletion of an evolutionarily conserved master regulator of cellulose-degrading enzymes does not result in compromised virulence during early infection. Instead, virulence is enhanced by compensatory induction of other virulence factors and subversion of plant immune responses. However, the pathogen is compromised during later stages of growth and reproduction because of its inability to metabolize cellulose as a carbon source. These findings prompt a different perspective on the importance of cellulose-degrading enzymes for fungal colonization of plants. —JM

Sci. Adv. 10.1126/sciadv.abl9734 (2022).

SOLAR CELLS

Organometallics stabilizing perovskites

Perovskite solar cells with inverted (p-i-n) structure can have greater stability and lifetimes than conventional n-i-p structures but usually have somewhat lower power conversion efficiencies (PCEs). Li *et al*. report that an organometallic compound, ferrocenyl-bis-thiophene-2-carboxylate, could stabilize a multication perovskite layer of an inverted perovskite solar cell. After 1500 hours of maximum power point operation, 98% of the 25.0% PCE was maintained. The solar cell also exhibited high stability in damp heat tests. —PDS

Science, abm8566, this issue p. 416

REVIEW SUMMARY

METAMATERIALS

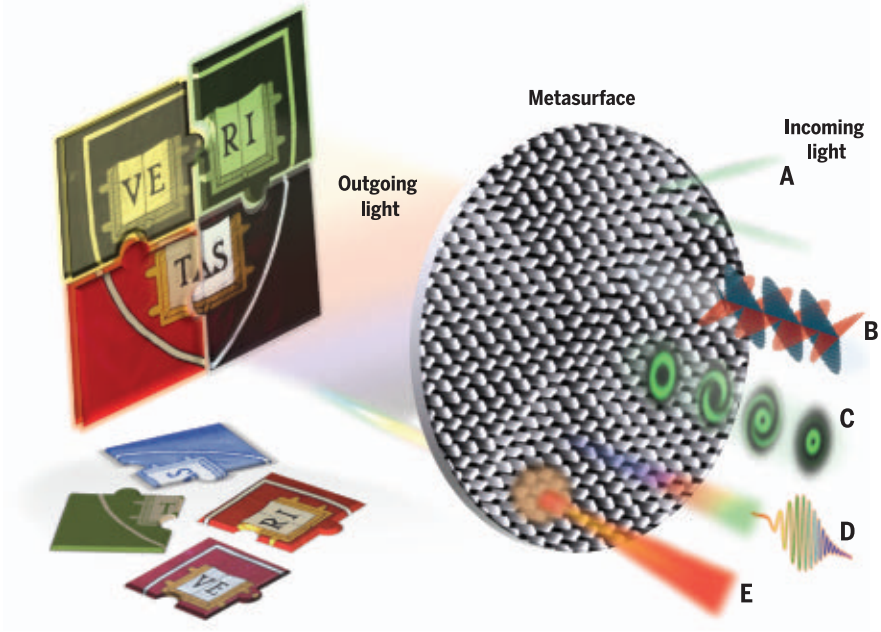
Tunable structured light with flat optics

Ahmed H. Dorrah* and Federico Capasso*

BACKGROUND: Structuring the degrees of freedom of light—including its phase, amplitude, and polarization—has opened new frontiers in science and technology alike. Adaptive cameras, microscopes, portable and wearable devices, optical communications, and laser machining are only a few of the domains that have evolved over the past decade owing to the advances in wavefront-shaping platforms. Flat optics composed of subwavelength-spaced optical scatterers—also known as metasurfaces or meta-optics—are key enabling tools for structured light not only for their compact footprint and complementary metal-oxide semiconductor (CMOS) compatibility but also because of their versatility and custom design. Although flat optics, or at least in its first generation, has led to the development of effects like anomalous refraction and diffraction-limited focusing, new classes of metasurfaces can now mold the flow of light in much more complex ways. Dispersion engineering and polarization optics are two prominent areas in which the metasurface's ability to spatially manipulate each wavelength and/or polarization state, independently, cannot be paralleled using bulk optical components. At the heart of these developments is a carefully engineered light-matter interaction at the level of the meta-atom, which allows a passive metasurface to produce this complex response.

ADVANCES: The ability to manipulate light in different ways, depending on its properties, is intriguing because it allows a passive device to produce many functions without the need for active switching—that is, light itself can be used as an optical control knob. For example, the same flat optic may behave as a lens or a mirror depending on the incidence angle of light. Likewise, by changing light's polarization, a metasurface can switch between different holograms or modify its focal length. In this spirit, a new generation of meta-optics can now perform parallel processing of the polarization of input light in the transverse plane or in 3D, reducing the function of many polarizers and waveplates into a single optical component that can be integrated in polarimeters and cellphone cameras. The spatial phase distribution of incoming light is another degree of freedom that can also be used as a switch, allowing a static flat optic to project different holograms by varying the helicity of the incident wavefront or its phase profile in general. Moreover, the ability to impart different phase and/or amplitude profiles on different wavelengths, independently, has enabled a wide class of versatile metalenses and compact pulse-shaping tools. Harnessing the nonlinear interaction of light with meta-atoms has also enabled multiwavelength holography on high harmonic-generated signals in addition to an asymmetric response. This versatility has made flat optics an ideal platform for the generation of structured light and has inspired many applications. The figure depicts the use of a metasurface as a multipurpose device (akin to a Swiss knife) that can mix and match output light by tuning the five above-mentioned control knobs, without the need for any complex circuitry to drive the meta-atoms. Tunable behavior of this kind relies on an intricate light-matter interaction at the nanoscale, which is often difficult to replicate with other wavefront-shaping tools. With such tunability, many existing technologies can be dramatically miniaturized, enabling compact spectrometers, polarization-sensitive cameras, lightweight augmented reality and virtual reality headsets, and biomedical devices.

OUTLOOK: As the area of structured light matures, the quest for more sophisticated metasurfaces is also on the rise, aided by advanced nanofabrication, powerful computation capabilities for revealing new meta-atom libraries, and recent developments in actively tunable materials for time-varying control. Structured light with tunable metasurfaces is poised to reveal new functionalities and to replace conventional optical systems with on-chip photonic components. This includes integrating metasurfaces in laser cavities, Fabry-Perot resonators, fiber-based devices, and active wavefront-shaping tools. With the emerging trends in inverse design and topology optimization, new standardized protocols for large-scale multilayer metasurface fabrication and innovative material platforms will push the limits of multifunctional meta-optics and structured light from 2D to 3D and from static to animate, thus tackling the open challenges in this wide field of research and unlocking many new paths. ■



Tunable structured light with static meta-optics. Different properties of input light may act as control knobs for tuning the optical response of the metasurface. These degrees of freedom include the angle of incidence (A), polarization state (B), orbital angular momentum (or spatial structure in general) (C), wavelength (D), and intensity level (manifested in a nonlinear interaction) (E). By changing one or more of these properties at the input of the metasurface, one can obtain a different light pattern at the output (depicted by the different puzzle pieces of Harvard University's logo). This tunability relies on an intricate light-matter interaction at the level of the meta-atom, which often cannot be replicated by other conventional wavefront shaping platforms.

Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA.
*Corresponding author. Email: dorrah@seas.harvard.edu (A.H.D.); capasso@seas.harvard.edu (F.C.)
Cite this article as A. H. Dorrah and F. Capasso, *Science* **376**, eabi6860 (2022). DOI: 10.1126/science.abi6860

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abi6860

REVIEW

METAMATERIALS

Tunable structured light with flat optics

Ahmed H. Dorrah* and Federico Capasso*

Flat optics has emerged as a key player in the area of structured light and its applications, owing to its subwavelength resolution, ease of integration, and compact footprint. Although its first generation has revolutionized conventional lenses and enabled anomalous refraction, new classes of meta-optics can now shape light and dark features of an optical field with an unprecedented level of complexity and multifunctionality. Here, we review these efforts with a focus on metasurfaces that use different properties of input light—angle of incidence and direction, polarization, phase distribution, wavelength, and nonlinear behavior—as optical knobs for tuning the output response. We discuss ongoing advances in this area as well as future challenges and prospects. These recent developments indicate that optically tunable flat optics is poised to advance adaptive camera systems, microscopes, holograms, and portable and wearable devices and may suggest new possibilities in optical communications and sensing.

Harnessing the properties of light defines how we experience the world around us. High-resolution microscopes, long-range telescopes, fast cameras, and spectrometers are a few tools that have shaped our understanding of the Universe, from the atomic to the astrophysical scale. At the heart of these developments lies an optical component that controls the properties of light as it interacts with matter. Several efforts have been made to manipulate light using artificial materials that do not naturally exist in bulk, inspired by nature's ability to structure light through diffraction, interference, and scattering—as observed, for example, in the colorful wings of butterflies (*1*). Early attempts toward this goal (from the fourth century) have relied on doping glass with metallic nanoparticles to modify its optical properties, producing intriguing structured coloration owing to the plasmonic resonance of the nanoparticles, as displayed in Roman artifacts and French cathedrals (*2*). Similarly, periodic multilayer stacks of alternating dielectrics in 1D (*3*) or 2D or 3D—that is, photonic crystals (*4*–*6*)—exhibit extreme reflection properties that deviate from their constituent materials. Engineering the optical response of matter by structuring its geometry or material composition provides new routes toward harnessing the properties of light in unconventional ways (*7*–*9*), similar to how semiconductor heterostructures have revolutionized optoelectronics. Metamaterials have emerged as new platforms for controlling light by virtually realizing anomalous values of effective permittivity and permeability (*10*, *11*), altering the output electromagnetic response on demand (*12*, *13*). Platforms of this nature are composed of periodic unit cells, referred to as meta-atoms, which are typically made of metallic or dielectric scatterers that are tightly packed in a lattice-like arrangement with subwavelength separation (*14*).

First introduced at microwave frequencies, the widespread use of metamaterials in optics was hampered by the challenges associated with their fabrication, a task that mandates 3D printing at the nanoscale. To surmount this obstacle, metasurfaces (also known as planar or flat optics or meta-optics) have evolved as a promising wavefront-shaping candidate thanks to their monolithic integration, compact design, and subwavelength control (*15*–*20*). Their earlier demonstrations in the optical regime have relaxed basic laws of physics (*21*–*24*) and introduced innovative mechanisms for wavefront tilting (*25*, *26*), holography (*27*–*29*), and diffraction-limited focusing (*30*, *31*). Notably, metasurfaces have gained wide popularity not only for their compact form factor and complementary metal-oxide semiconductor (CMOS) compatibility but also because of their unprecedented control of polarization (*32*, *33*) and dispersion engineering (*34*). The former has enabled point-by-point polarization transformations at the nanoscale (*35*, *36*), whereas the latter has allowed broadband focusing and multiwavelength holography (*37*). These are nontrivial advances that, to date, cannot be replicated with conventional polarization optics or other wavefront-shaping tools. As this area of research matured, the complexity of metasurface design also progressed, enabling more sophisticated functions and a tunable output response. For instance, a wide class of active metasurfaces can produce time-varying behavior that can be precisely controlled with external stimuli, bringing new physics and applications (*38*–*40*). Notably, static metasurfaces, with the proper design, can also change their output response by tuning the properties of input light (*41*). Versatile devices of this kind can realize fast all-optical switching without the need for electronic circuitry to modulate the output response, thereby saving footprint, complexity, and cost.

In this review, we take a closer look at these static devices, focusing on passive meta-optics that can tune their output behavior in response to changing one or more degrees of freedom of input light. Devices of this nature typically rely on an intricate light-matter interaction at the meta-atom level, which cannot be easily replicated by other platforms (like, for example, spatial light modulators). This tunability is often realized by different types of resonances (Mie scattering, Fano resonances, bound states in the continuum) (*42*) or shape birefringent meta-atoms (*35*, *36*) and free-form topologies (*43*, *44*), thus adding to the existing structured-light toolkit enabled by digital holography (*45*). We discuss possible technological consequences of these metasurface-based devices, their anticipated challenges, and open areas for innovation. Besides structuring light, we also highlight new methods for structuring the dark (namely, phase singularities) with flat optics and hint to its applications.

Angle-dependent and directional response

Angle dependence is usually an undesired feature in conventional lenses, holograms, and beam-steering applications because it manifests in the form of diffractive loss, distortion, or coma aberration. Independent control of each input wave vector can mitigate these effects and enable new functions. However, achieving this task is not straightforward because, on a macroscopic scale, metasurfaces (similar to Fresnel lenses) resemble diffraction gratings, which are angle sensitive, whereas on a microscopic level, the response of simple meta-atom geometries (like cylindrical nanopillars) is typically maintained over a wide range of angles. Local and nonlocal metasurfaces tackle this dilemma in conceptually different ways. The former acts on incident light, point by point, in real space (e.g., gradient metasurfaces), whereas the latter relies on neighbor-to-neighbor meta-atom interactions and often uses sharp resonances to collectively modify the output response in momentum space (akin to photonic crystal slabs) (*46*). Early examples of local metasurfaces, with angle dependence, used U-shaped meta-atoms made of amorphous silicon supported on a reflective aluminum mirror with a thin spacer of silicon dioxide in between (*47*). In this configuration, the meta-atoms act as a multimode resonator, imparting different phase delays on a discrete set of incidence angles. An angle-multiplexed metasurface that generates two distinct holograms for 0° and 30° incident light (wavelength λ = 915 nm) was thus demonstrated. Another

Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA.
*Corresponding author. Email: dorrah@seas.harvard.edu (A.H.D.); capasso@seas.harvard.edu (F.C.)

approach exploits the coupling between neighbor meta-atoms to create multifunctional devices that, depending on the excitation angle, may project a different hologram with phase (*48*) or complex-amplitude modulation in the visible range (*49*) and can act as a mirror or a half-wave plate in the near-infrared (NIR) range (*50*). Figure 1A shows a configuration that relies on mutual coupling between meta-atoms (comprising metal-insulator-metal) to achieve an angle-dependent response (*49*), enabling different holograms to be generated in phase and amplitude by changing the angle of incidence (Fig. 1B).

Additionally, dielectric metasurfaces that combine interleaved superpixels with the concept of detour phase have been used to expand the number of angle-dependent functions: for instance, generating four distinct holograms in response to four distinct angles of incidence (*51*). The coupling between meta-atoms is highly sensitive to the surrounding environment, suggesting unexplored mechanisms for sensing. For instance, a nonlocal germanium-based metasurface has been used to realize highly sensitive and sharp resonances, with spectral position controlled by the incidence angle of mid-infrared (IR) light (Fig. 1C) (*52*). This angle-multiplexed approach delivers more than 200 resonances for angles between 13° and 60° with a spectral coverage between 1100 and 1800 cm^{-1} and a spectral resolution below 5 cm^{-1}. Analyte molecules that match the angular position of the

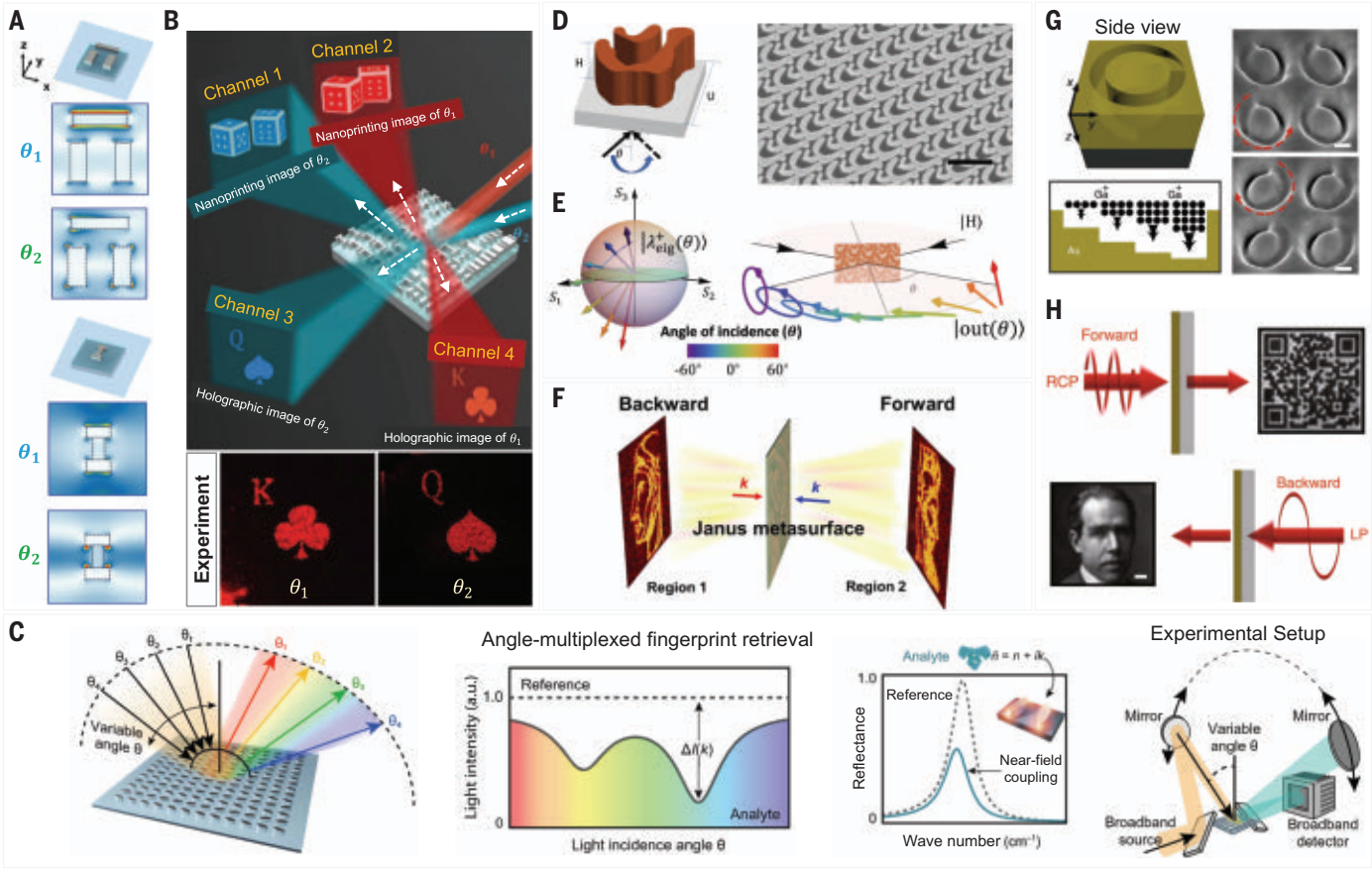


Fig. 1. Angle-dependent and directional metasurfaces. (**A**) Coupled metal-insulator-metal meta-atom geometries in which the electric energy density distribution depends on the illumination angle, θ. Reproduced with permission from (*49*). (**B**) A metasurface composed of these unit cells can impart different functions for different θ, such as "nanoprinting" images (channels 1 and 2) and holographic images (channels 3 and 4) (top). Measured data for two independent complex-amplitude holograms (at λ = 633 nm) for θ = 0° and 30° (bottom). Reproduced with permission from (*49*). (**C**) Germanium-based dielectric metasurface with a high quality factor produces sharp resonances with distinct resonance frequency for each incidence angle over a broad range. Strong near-field coupling between the dielectric resonators and the molecular vibrations of a nearby analyte induces a pronounced attenuation of the resonance line shape that is associated with the vibrational absorption bands. This configuration is suited for surface-enhanced mid-IR molecular absorption spectroscopy. A simple implementation is depicted on the far right. a.u., arbitrary units; $\Delta I(k)$, normalized intensity as a function of the imaginary part of the complex refractive index. Images reproduced from (*52*). (**D**) Free-form amorphous silicon nanostructures patterned on top of a glass substrate exhibit an angle-dependent polarization response at 1550 nm (left). Here, the nanostructure's height (H) = 1500 nm and center-to-center separation (U) = 600 nm. A SEM image of a fabricated sample is shown on the right. Scale bar is 1 μm. Images reproduced from (*44*). (**E**) On the left, the arrows represent the angle-dependent birefringence axis visualized in polarization space. Different colors refer to different angles of incidence. By varying the angle, the device can be continuously tuned between linear and elliptical birefringence (right). Images reproduced from (*44*). *S*, Stokes components. (**F**) A directional metasurface can project one face of Roman god Janus while totally concealing the other, depending on the direction of input light. *k*, wave vector. Reproduced with permission from (*63*). (**G**) Janus metasurfaces can be implemented by making use of plasmonic helical meta-atoms with an asymmetric chiral response. Using focused ion beam milling with different doses of gallium, the depth of the spiral groove can be adiabatically increased along the red dashed arrows, as shown in the SEM images. Scale bars are 100 nm. Images reproduced from (*62*). (**H**) Direction-controlled polarization-encrypted data storage. Right-hand circularly polarized (RCP) light at 800 nm can generate a specific QR code in forward transmission, whereas linearly polarized (LP) light creates a distinct image in reverse. Scale bar is 10 μm. Images reproduced from (*62*).

resonances incur strong modulation of the far-field response through surface-enhanced near-field effects. By detecting the angle-resolved reflectance signal from such metasurfaces before and after coating with the analyte molecules, the full spectral content of the molecular absorption fingerprint can be retrieved. Another approach uses off-axis multiplexed holograms to project distinct images on a discrete set of incidence angles (*53*). Note that obtaining a continuous angle-tunable function is a more daunting task because it mandates controlling the phase derivative with respect to the incident angle. Complex functions of this kind can be achieved using free-form metasurfaces. For example, Fig. 1D shows a topology-optimized polarization meta-optic that modifies its birefringence depending on the incidence angle (*44*). Light incident on this device would witness linear, circular, or elliptical birefringence depending on its angle (Fig. 1E). Inverse-designed devices of this type use the nonlocal interaction between adjacent meta-atoms to directly operate on incident light in momentum space. Besides an angle-dependent response, nonlocal metasurfaces have enabled many other complex functions such as performing mathematical operations (*54*, *55*), all-optical analog image processing and edge detection (*56*, *57*), and space compression (*58*, *59*), as well as externally tuned laser cavities (*60*). The latter relies on supercell metasurfaces whose diffraction orders can be efficiently and independently controlled at extreme angles.

Besides the angle of incidence, the direction of propagation (forward or backward) has been exploited for tuning the response of static metasurfaces. Commonly referred to as Janus metasurfaces, these directional devices exhibit different coloration (*61*) or asymmetric transmission (*62*, *63*) by reversing the direction of illumination from the front to the back side. Figure 1F shows a schematic of a metasurface that projects one face of Roman god Janus while totally concealing the other depending on the direction of input light (*63*). This has been realized using meta-atoms made of cascaded subwavelength anisotropic impedance sheets. By introducing a gradual rotational in each sheet, linearly polarized light will undergo asymmetric transmission. Similar behavior can be achieved with helical meta-atoms (Fig. 1G), which exhibit circular dichroism as large as 0.72 with forward transmission, owing to spin-dependent mode coupling, while incurring giant linear dichroism up to 0.87 in reverse, with high selectivity for the azimuthal angle of linearly polarized light (*62*). Despite their 3D chiral geometry, these meta-atoms are fabricated with single-step focused ion beam milling. By creating a metasurface with two meta-atom enantiomers of specific rotation angles, direction-controlled polarization-encrypted holography can be realized. For instance, Fig. 1H shows a scenario in which a binary QR code image is displayed in the forward direction under right-hand circularly polarized light, whereas a distinct grayscale image is generated in the backward direction under linear polarization. This scheme can be useful in data encryption, optical storage, and information processing. To this end, angle dependence and directionality can be combined to generate distinct holograms in reflection and transmission simultaneously (*64*, *65*).

Polarization-switchable behavior

Unlike conventional bulk polarizers and waveplates, which are components that control light's polarization globally, the metasurface counterparts enable point-by-point polarization transformations at the nanoscale because of the shape-dependent birefringence of their meta-atoms. Consider, for instance, the rectangular nanofin shown in Fig. 2A. Light transmitted through this "nanoantenna" will experience two independent phase delays along the major and minor axes, depending on the dimensions of the nanofin. Additionally, the nanofin's angular orientation—that is, its birefringence axis—can be rotated, thereby offering a third degree of freedom that typically manifests as a Pancharatnam-Berry phase (*66*). Replicating this scheme with other wavefront-shaping methods such as spatial light modulators (which only operate on one polarization at a time) is challenging because it requires multiple interactions with incident light or cascading two or more devices (*36*, *67*). Metasurfaces thus stand out as a powerful tool for polarization control and vector-beam generation (*35*, *36*). Earlier attempts, besides polarization gratings (*68*, *69*), have relied on subwavelength Pancharatnam-Berry microstripe gratings at a wavelength of 10.6 μm for wavefront tilting (*26*) and vector-beam generation (*70*) and have exploited plasmonic metasurfaces for manipulating light's chirality (*71*, *72*). In parallel with these efforts, exotic classes of structured light have been created through spin-orbit coupling (*73*). Liquid crystal q-plates are one popular manifestation of the latter in which conjugate pairs of vortex beams—donut-shaped structured light with helical phase profiles, on-axis singularity, and orbital angular momentum (OAM) (*74*)—are generated while reversing the chirality of incident circularly polarized light (*75*). Another approach based on shared aperture antenna arrays and asymmetric harmonic resonances from metal-insulator-metal meta-atoms has enabled spin-controlled structured light (*76*). Reflective-type plasmonic nanoantennae have also been used to generate polarization-dependent dual images over a broad band in the visible range (*77*). In addition, dielectric metasurfaces made of elliptically shaped meta-atoms have been used to construct independent pairs of vectorial modes (radially and azimuthally polarized light) in response to input x and y linearly polarized light in the visible range (*32*, *78*) and have enabled spatial-mode multiplexing at the telecom (1550 nm) wavelength, using linear polarization as a switch (*79*, *80*). More generally, other states of polarization such as circular and elliptical have been used to switch between two independent phase holograms (*33*) or two different vortex beams of arbitrary topological charge (helicity), enabling arbitrary spin-orbit coupling (*81*); the latter device is known as a J-plate. There are now numerous examples of polarization control with metasurfaces, from polarization converters to polarization-dependent holograms, varifocal lenses, nanoprinting, and vortex-beam generators, which are comprehensively reviewed in (*35*, *36*).

We highlight more recent meta-optics with complex polarization-switchable behavior; one example of this is a metasurface that generates a far-field profile that performs parallel polarization analysis of incident light (Fig. 2B) (*82*). The intensity of each polarization state (denoted by the green arrows) obeys Malus's law, that is, the intensity is proportional to the projection (dot product) of the incident polarization onto that particular state. Hence, this class of metasurfaces, dubbed Jones matrix holograms, enables visual full-Stokes polarimetry of input light by pictorially reading the projected pattern. Figure 2C shows another metasurface that makes use of superpixels composed of four rectangular nanofins to project two complex-amplitude holograms in response to any input pair of orthogonal polarizations (*83*, *84*). Although this multifunctional behavior is limited to two orthogonal polarization channels, chirality-assisted phase modulation deploys cascaded metasurface layers to decouple all four components of the Jones matrix that describe each meta-atom (*85*). Using this approach, four distinct wavefronts are encoded on the input-output polarization channels L-L, L-R, R-L, and R-R (L-R denotes left- and right-handed circular polarization at the input and output of the device, respectively). Figure 2D shows a metadeflector that uses this concept, steering an input wavefront to four different directions by decoupling all input-output co-polarized and cross-polarized channels (*85*). Similarly, by sandwiching a birefringent metasurface between two polarizers, seven different images have been encrypted on different input-output polarization channels at an 800-nm wavelength (*86*). Multichannel holography has also been enhanced by including wavelength as an additional degree of freedom (*87*). Notably, parallel polarization processing and analysis may devise new techniques for polarization characterization. For instance, full Stokes polarimetry [which requires at least four intensity measurements of input light onto four different analyzers (*88*)] can be replaced with a single metasurface

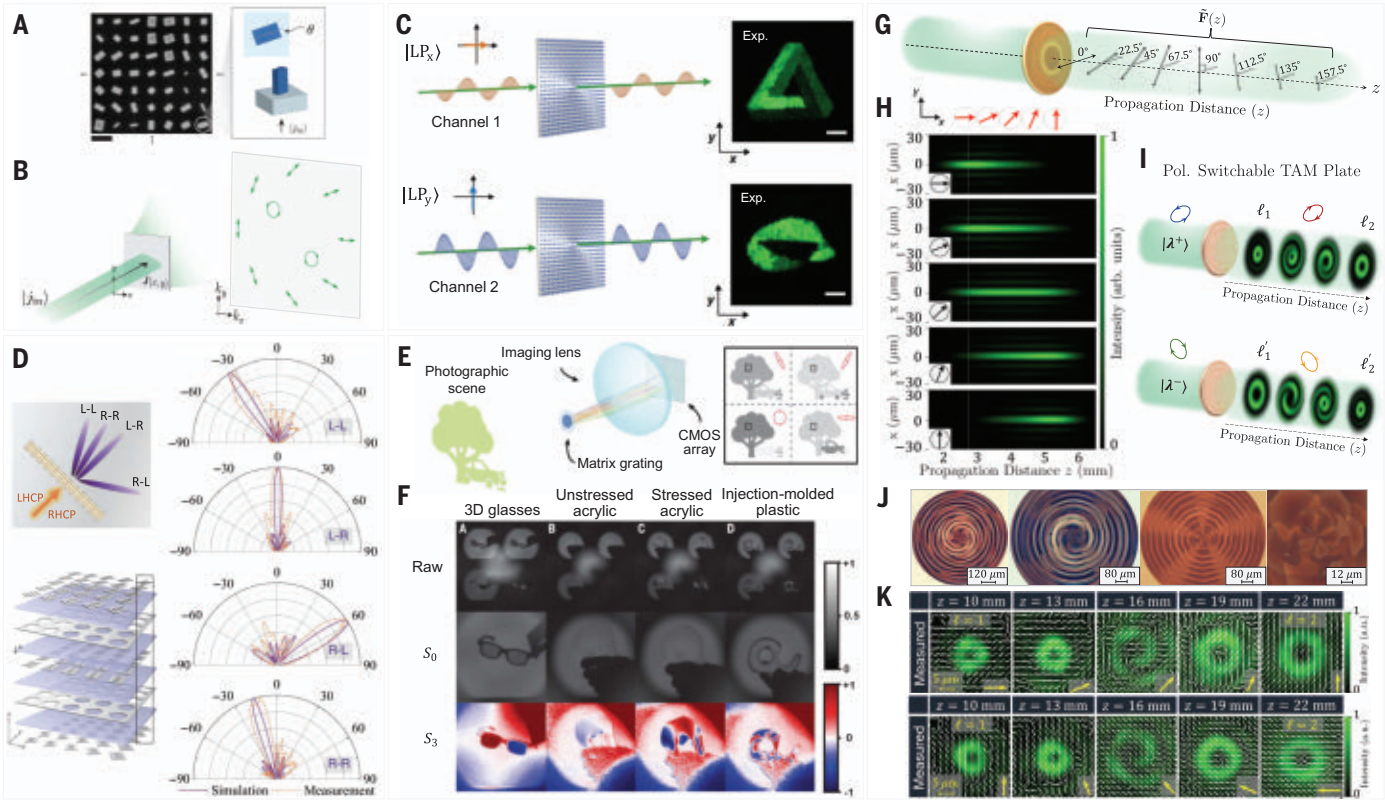


Fig. 2. Metasurface with polarization-dependent response. (**A**) A birefringent metasurface composed of dielectric nanofins. SEM image of a small section is shown. Scale bar is 500 nm. Images reproduced from (*82*). (**B**) Polarization-analyzing hologram. Under coherent illumination $\boldsymbol{j}_{in}$, light will be projected into specific far-field pixels depending on the input polarization. Each drawing on the screen acts as an "analyzer" that will exhibit an intensity level for its shown polarization state (depicted by the green arrows) in accordance with Malus's law. J, spatially varying Jones matrix. Images reproduced from (*82*). (**C**) Polarization-dependent complex-amplitude holograms. Far-field images of a "Penrose triangle" (top) and "Möbius strip" (bottom) are generated (at λ = 530 nm) in response to x- and y-polarized input light, respectively. Scale bars are 200 μm. LP, linear polarization. Images reproduced from (*83*). (**D**) Chirality-assisted metadeflector can steer the input beam into different directions under four input-output polarization channels of circularly polarized light: L-L, L-R, R-L, and R-R (from top to bottom). Here, L-L denotes input LCP–output LCP light. The device consists of five metallic layers and four dielectric substrates. Simulated and measured normalized far-field patterns are shown at 10 GHz. LHCP, left-handed circular polarization; RHCP, right-handed circular polarization. Images reproduced from (*85*). (**E**) Matrix gratings can analyze for many polarization states in parallel at the far field. Incident light will distribute its energy onto the four orders subject to Malus's law. Integrating this meta-optic with a conventional CMOS sensor enables real-time polarization imaging with no moving parts. Images reproduced from (*89*). (**F**) Polarization imaging makes use of the measured S_3 Stokes component. In each example, raw sensor acquisition, S_0 (intensity image), and S_3 are shown. Examples include 3D glasses with opposite circular polarizers and a laser-cut acrylic piece that is stressed by hand-squeezing, displaying stress birefringence in its S_3 image. The black-to-white scale bar represents normalized intensity; the blue-to-red scale bar represents the value of of the chiral Stokes component. Images reproduced from (*89*). (**G**) Longitudinally variable polarization optics perform many polarization operations, simultaneously, along the optical path. The black arrows depict the virtual principal axis orientation of the polarizing element at each z plane. Images reproduced from (*99*). $\tilde{F}$, target polarization transformation. (**H**) Longitudinal profiles of the generated "pencil-like" beam for each incident polarization that is depicted in the inset. The on-axis intensity distribution continuously shifts away from the source by rotating the input polarization from 0° to 90°. Images reproduced from (*99*). (**I**) Polarization-switchable TAM plates enable arbitrary spin-orbit coupling in 3D, mapping any pair of orthogonal polarizations ($|\lambda^+\rangle$ and $|\lambda^-\rangle$) into two propagation-dependent vortices with varying OAM state, ℓ. Images reproduced from (*100*). (**J**) Optical micrographs of the devices. Images reproduced from (*100*). (**K**) Versatile TAM plates can control light's polarization and OAM along the optical path. The arrows depict the evolution of output polarization under x-polarized (top) and y-polarized (bottom) illumination. Images reproduced from (*100*).

(Fig. 2E) (*89*). This metagrating projects input light onto four diffraction orders, each analyzing for a different polarization in the far field. The underlying principle relies on matrix Fourier optics, which recasts plane-wave decomposition in a two-by-two matrix form, allowing each plane-wave term in the sum to be weighted by a Jones matrix (as opposed to a complex scalar) coefficient (*89*). This formalism generalizes a wide body of work that seeks to generate vectorial structured light under the assumption of a fixed input polarization (*90*–*95*) by lifting that constraint. It also mitigates unwanted diffractive losses and limited functionality of interlaced metagratings (*18*, *96*–*98*). By interfacing matrix gratings with a commercial CMOS sensor, the spatial polarization profile of a scene can be imaged in real time, enabling a polarization-sensitive camera that provides extra information compared with raw intensity images (Fig. 2F).

Besides controlling polarization in the transverse plane (in 2D), a new class of meta-optics has enabled parallel polarization transformations along the optical path. In this case, a single meta-optic can mimic an arrangement of many polarization optics cascaded in series (Fig. 2G) (*99*). Light incident on this device is converted into a quasi-nondiffracting (pencil-like) beam that changes its polarization state as if encountering different polarizers and

waveplates at each z plane thereafter. For example, Fig. 2H exhibits the response of a meta-optic that analyzes for a set of linear polarizations between 0° and 90° along the z direction, as depicted by the red arrows. By rotating the input linear polarization from 0° to 90°, the centroid of the output beam is continuously translated along the axial direction, following the projection on the red arrows and in accordance with Malus's law (*99*). This polarization-switchable axicon may find use in optical trapping applications in which rapid polarization modulation at the input can modify the output scattering forces and radiation pressure on demand. Besides polarization, this design methodology can be generalized to control light's OAM along the propagation direction. Figure 2I shows the response of total angular momentum (TAM) plates (*100*). By switching the input polarization between any two specified orthogonal bases, the device produces two distinct optical vortices in which the OAM state ℓ (wavefront helicity) varies with propagation, as signified by the change in the beam's diameter. The optical micrographs of these devices (Fig. 2J) express very characteristic patterns. Simultaneous and independent control over both components of angular momentum (spin and orbit) has also been achieved (Fig. 2K) (*100*). In this case, the vortex beam changes its OAM state and rotates its polarization state adiabatically as a function of z, regardless of the incident polarization. Notably, the evolution in angular momentum occurs only locally over the central region of the beam because of spatial multimode beating (i.e., controlled interference between copropagating modes with different wave vectors). The total OAM is conserved when integrated across the entire transverse plane. 3D structured light can be used in refractive index sensing (*101*, *102*), free-space optical communications (*103*), and micromanipulation at multiple planes (*104*).

Tuning with structured light

Because a photon's spin is constrained to two degrees of freedom, polarization-switchable metasurfaces are primarily limited to mapping two orthogonal polarizations to two output wavefront profiles. Multichannel holography (*86*, *87*) can relax this constraint by decoupling different input-output polarization channels, but with additional bulk optics and an upper limit on the number of generated holograms, as described in the previous section. To overcome these challenges, OAM holography with metasurfaces has emerged as a versatile wavefront-shaping tool (*105*–*107*). It relies on mapping an orthogonal set of vortex beams (modes with helical phasefront) to an arbitrarily large number of output holograms (constrained by the aperture size). Figure 3, A to C, depicts three variations: OAM-conserving, -selective, and -multiplexing meta-holograms (*105*). OAM-conserving holograms produce pixelated images while preserving the OAM property of incident OAM beams in each pixel of the reconstructed image (Fig. 3A). To achieve this, the holographic image is spatially sampled by an OAM-dependent 2D Dirac comb function to avoid spatial overlap of the helical wavefront kernel at the image plane (upon its convolution with the hologram). OAM-selective holograms, on the other hand, are sensitive to the input OAM mode because they project a holographic image in response to a particular helical mode. This is realized by adding a spiral phase $e^{-i\ell\phi}$ on an OAM-conserving meta-hologram, where e is Euler's number, i is the imaginary unit, and ϕ is the azimuthal phase. In this case, the target image will only be projected if the conjugate mode $e^{i\ell\phi}$ is incident on the metasurface. Notably, adding multiple OAM-selective holograms (designed to be sensitive to different OAM numbers) on the same metasurface will create an OAM-multiplexing meta-hologram that maps different OAM states to distinct holographic images (Fig. 3C). Gallium nitride (GaN) metasurfaces were used to demonstrate these three types of holograms in the visible range (632 nm), by mapping four helical modes to four different holograms (*105*).

Furthermore, owing to their orthogonality and unbounded size, many OAM modes can be multiplexed by a single meta-hologram while maintaining high spatial resolution. For example, OAM states ranging from an ℓ of −50 to 50 impinging on an OAM-multiplexing meta-hologram can sequentially address 200 OAM-dependent orthogonal holographic image frames (100 images at two different axial positions, z), allowing an all-optical holographic video display without any mechanical scanning (Fig. 3D) (*106*). Here, the metasurface is made of a 3D-printed polymer and can achieve complex-amplitude modulation at 633 nm. Notably, more than one hologram can be encoded on the same OAM state by using the input-output polarization channel as an additional degree of freedom (Fig. 3E) (*108*).

Besides OAM, more general wavefront distributions can serve as control knobs to project distinct holographic images (Fig. 3F) (*109*). By illuminating a silicon metasurface with a judiciously engineered wavefront at 633 nm, the encrypted image can be displayed, whereas under uniform illumination, random noise is generated. This approach relies on dividing the target phase profile of the holographic image into two distributions that are imposed on the metasurface and the illumination beam. Even though the metasurface is static, the incident beam contains a large parameter space that can alter (reprogram) the output beam, suggesting new techniques for information security and authentication. Similarly, an all-optical solution for secret key sharing has been proposed using cascaded meta-surface holography (*110*). This configuration can be used to split and share encrypted holographic information across multiple metasurface layers (Fig. 3G). A set of metasurfaces are designated as "shares." Each of them contains an encoded phase-only Fourier hologram, which can be reconstructed in the far field upon illumination with circularly polarized light, serving as specific identifiers for each metasurface. Meanwhile, when two metasurfaces are stacked 100 μm apart, the illumination of the cascaded configuration creates a new holographic image that is distinct from the two single-layer holograms. This cascaded holographic image can be used as an optical secret that is only revealed if both metasurfaces are stacked and can be extended to a larger set of shares. The concept relies on the fact that the same cascaded phase mask can be built up through combinations of different single phase masks without revealing any information about the shared secret within the single-layer images. It has been implemented with an iterative gradient optimization approach that uses the "automatic differentiation" feature of machine learning. In the future, a reprogrammable version of this scheme can be realized by combining a metasurface as the main share with a digitally addressed spatial light modulator as the deputy share. In addition to structured light, complex patterns of structured dark can potentially be used as metasurface knobs. For example, beyond the 1D singularity carried by OAM modes, 2D singularity sheets have been engineered by minimizing the real and imaginary components of the field over an arbitrarily chosen geometry, as shown in Fig. 3H (*111*). By maximizing the phase gradient normal to the region of interest, heart-shaped singularity sheets have been generated. Similarly, vectorial singularity sheets, over which the polarization state is undefined, have also be constructed. These degrees of freedom provide additional knobs for meta-holograms and may suggest new sensing schemes owing to their extreme sensitivity to perturbation (*112*).

Multiwavelength control

Dispersion, or wavelength dependence, is a profound property of all optical materials. It imposes a fundamental limit on achromatic focusing, broadband holography, and high-rate data transmission. Decades of effort have attempted to tailor dispersion by tuning the chemical composition (refractive index) of matter, which is often a daunting and nonscalable task. Metasurfaces provide a more flexible route to dispersion engineering through shaping the geometry of meta-atoms, thereby altering their effective refractive index in a reproducible manner at the nanoscale. Dispersion-engineered metasurfaces can emulate the dispersion properties of refractive or diffractive optics, on demand, and have subsequently emerged as a

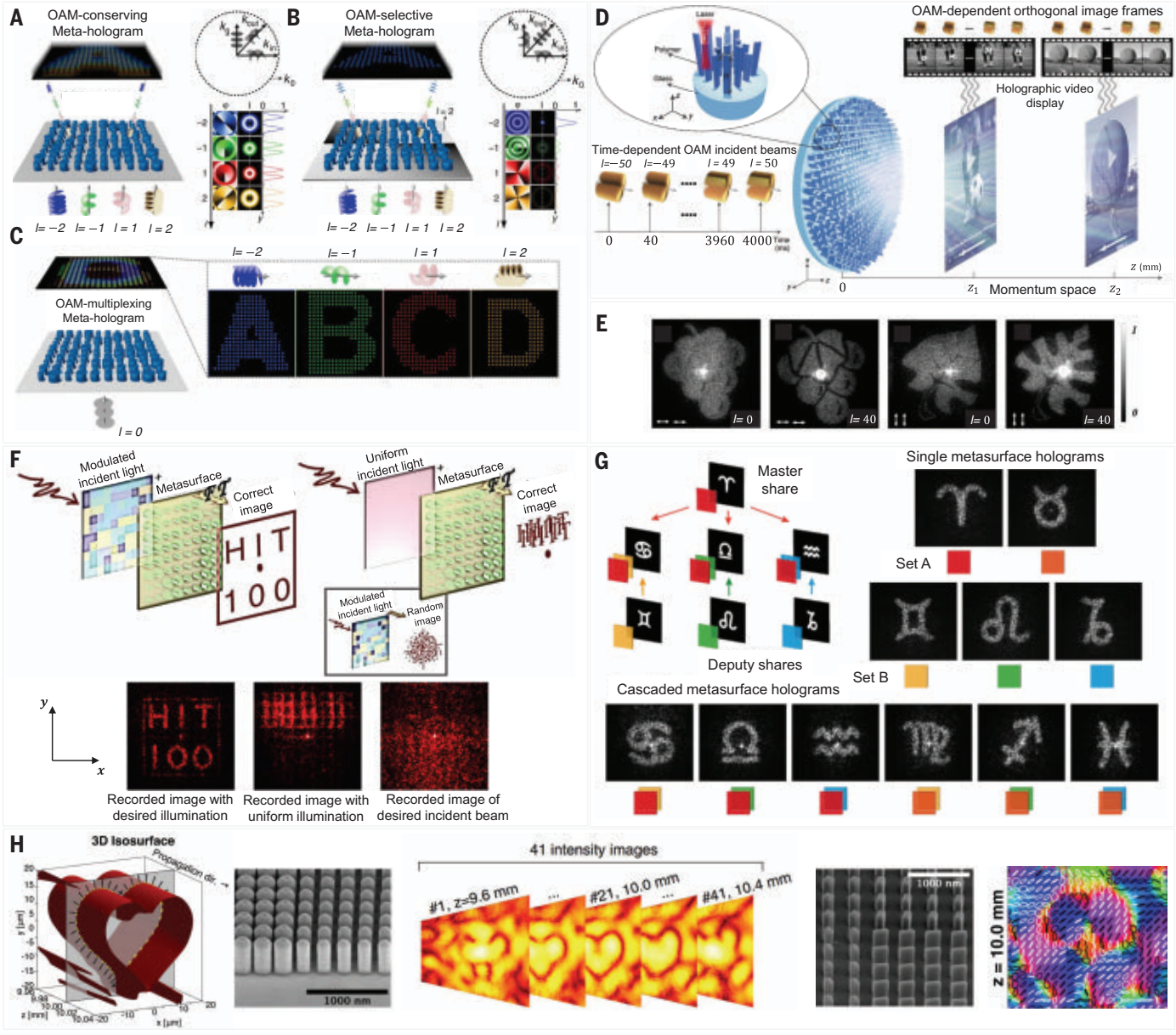


Fig. 3. Structured light as a metasurface knob. (**A**) OAM-conserving holograms can project a pixelated image where each pixel carries the same OAM as the source, as depicted by the arrows in the spatial frequency (*k*-space) domain (top right). The corresponding phase (ϕ) and intensity (*I*) distributions of single pixels in the reconstructed images are shown at the bottom right. Pseudocolors visualize different OAM modes. Images reproduced from (*105*). (**B**) OAM-selective holograms can project an image in response to a specific input OAM mode. The phase and intensity of each pixel in the reconstructed holographic images are shown on the right. High intensity is achieved in each pixel if the incident OAM mode matches the specified design ($\ell = -2$). Images reproduced from (*105*). (**C**) OAM-multiplexing holograms can reconstruct many OAM-dependent holographic images. Images reproduced from (*105*). (**D**) OAM holography can realize time-varying image frames by mapping a range of input OAM states (from $\ell = -50$ to 50) to output images, creating a movie. This can occur at multiple *z* planes along the optical axis. Here, the metasurface is fabricated using 3D laser printing, as depicted in the inset. The nanofin's height (*H*) and rotation (θ) provide independent control of amplitude and phase. Reprinted by permission from Springer Nature (*106*). (**E**) Polarization-encrypted OAM holography makes use of the input-output polarization channels (white arrows) to encode additional images on each OAM state. The black-to-white scale bar represents normalized intensity. Reproduced with permission from (*108*). (**F**) Encrypted meta-holograms project the desired image only when illuminated with the correct light structure and generate random noise otherwise. Images reproduced from (*109*). (**G**) Secret key sharing using "master share" metasurface holograms (set A) and three "deputy share" metasurface holograms (set B). Every metasurface in each set can project a distinct Fourier hologram. By stacking two metasurfaces from sets A and B, a new holographic image is generated. Each metasurface is represented by a specific color. Images reproduced from (*110*). (**H**) Heart-shaped optical singularities can be engineered by creating an isosurface of low field intensity. The phase gradient is maximized in the directions indicated by the arrows (left). A titanium dioxide metasurface made of cylindrical nanofins can generate the heart-shaped dark patterns (middle), whereas a birefringent metasurface composed of rectangular nanofins can create vectorial singularity sheets with undefined polarization azimuth (far right), as indicated by the black and white lines. Images reproduced from (*111*).

promising platform for achromatic focusing, wavelength-dependent holography, multifunctional devices, and pulse shaping, as comprehensively reviewed in (*20*, *34*, *37*). Note that achromaticity is dominated by the phase and group delay (first-order dispersion) of the phase shifters. For example, to maintain a broad-band performance, the phase shifters should exhibit smooth dispersive responses by avoiding sharp resonances. On the other hand, operating near sharp resonances can effectively decouple the phases at different wavelengths, allowing a multifunctional response.

Multiwavelength control has been realized with interleaved superpixels, guided-mode resonances, coupled meta-atoms, and stacked metasurfaces, among others. Figure 4A shows one implementation of the former in which a dielectric metasurface manipulates the phase response of red, green, and blue wavelengths independently (*113*). The meta-molecule consists of three kinds of silicon nanoblocks, each imparting a geometric phase on a specific wavelength by varying the in-plane orientation of each waveplate-like nanofin. Despite its straightforward implementation, spatial interleaving imposes an upper limit on efficiency and introduces undesired meta-atom coupling, which degrades the image quality, producing ghost images and unwanted diffraction orders. To overcome these limitations, vertically stacked metasurfaces have been proposed where each layer is optimized for a particular incoming wavelength, enabling achromatic focusing and a multiwavelength response, albeit at the expense of complex fabrication requirements (*114*). These challenges can be mitigated by using dispersion-engineered metasurfaces such as the one depicted in Fig. 4B (*115*). This scheme makes use of a reflective metasurface to effectively double the propagation phase per pillar compared with operating in transmission. It also introduces wavelength-dependent guided-mode resonances, which arise when incident light couples to the leaky surface modes and reradiates into free space through phase matching. This creates rapid phase gradients around the resonances, enabling broad phase coverage while decoupling the phases imparted on each wavelength. Using this approach, red, green, and blue wavelengths have been mapped to three different phase profiles (Fig. 4B). Bilayer metasurfaces can further expand the design space, enabling wavelength-selective holography with complex-amplitude modulation, as depicted in Fig. 4C (*116*).

Metasurface-assisted configurations have also been used in pulse shaping and spatiotemporal light control. For example, Fig. 4D depicts a frequency-gradient metasurface created by combining a frequency comb source (centered at λ = 720 nm) with a passive metasurface (made of silicon nanopillars on sapphire substrate) to achieve ultrafast dynamic beam steering without any mechanical components (*41*). The underlying concept relies on mapping each incoming spectral line into a spatial optical mode with a different wave vector. Because the spectral lines are phase-locked, their individual spatial patterns constructively interfere to generate a 4D optical pattern in which the spatial light intensity distribution naturally evolves in time (steering 25° in 8 ps).

Figure 4E shows another metasurface-based setup for 1D pulse shaping, realized by placing a metasurface made of polycrystalline silicon nanopillars at the focal plane of a Fourier-transform—that is, spectral dispersing-recombining—pulse synthesizer (*117*). The metasurface provides independent phase and amplitude control for each spectral component of a 10-fs pulse centered at 800 nm, thereby tailoring its temporal characteristics at will. More compact setups for pulse shaping, without the use of diffraction gratings, can be achieved through direct interaction with a single metasurface (Fig. 4F) (*118*). In this case, a broadband pulse compressor nanocoating compensates for the group delay dispersion of up to 2-mm-thick fused silica glass in the visible-to-NIR spectral region with a bandwidth of up to 80 nm. The group delay characteristics are determined by geometric properties of the nanopillars rather than their material dispersion, as shown in Fig. 4F, providing a versatile approach that can be adapted to different spectral regions and applications.

Nonlinear metasurfaces

The use of light's degrees of freedom as control knobs can naturally be extended to harness nonlinear interactions. Note that obtaining a strong nonlinear response from optically thin structures requires much stronger light-matter interactions than natural bulk nonlinear media. Nonlinear metasurfaces tackle this dilemma by judiciously structuring the shape and size of the nanoscatterers, thereby enhancing the nonlinear effects over very small volumes by relaxing phase-matching and symmetry rules (*119*, *120*). Earlier work relied on the excitation of localized surface plasmon polariton resonances to strongly enhance the electromagnetic field in the vicinity of the meta-atoms—an effect that can be tailored by designing the meta-atom geometries. Configurations of this type often use metal-dielectric interfaces where the nonlinearities stem from the asymmetry of the potential, confining the electrons at the material interface (*121*). For instance, Fig. 5A shows a meta-atom made of a split-ring resonator, which provides strong anisotropic response in the linear regime as well as highly efficient second-harmonic generation (SHG) (*122*). Circularly polarized light with spin $-\sigma$ interacting with this meta-atom (of angular orientation ϕ) will accumulate a geometric phase of $2\sigma\phi$ (σ is ±1 depending on the input polarization handedness), whereas the SHG signals with spin σ and $-\sigma$ will acquire a Berry phase of $\sigma\phi$ and $3\sigma\phi$, respectively. Thus, by designing orientation angles of each meta-atom, three different spatial phase distributions can be imparted on the fundamental, co- and cross-polarized SHG under left-handed circular polarization (LCP), as shown in Fig. 5B. Here, we refer to the Berry phase acquired by the SHG signal as the nonlinear Berry phase. Moreover, nonlinear metasurfaces based on multilayered V-shaped gold nanoantennas have enabled polarization-multiplexed holography on the third harmonic–generated (THG) signal (λ = 422 nm), making it possible to project different images at multiple propagation distances by changing the incoming polarization (*123*).

Recently, all-dielectric metasurfaces have received much attention because of their low optical loss, strong field overlap, and systematic control over their dispersion properties, as reviewed in (*120*). Here, we highlight one application that made use of a silicon metalens to enhance the nonlinear conversion efficiency for the THG process (*124*). Figure 5C depicts the proposed nonlinear lens scheme. Here, the fundamental wave (1550 nm) incorporates the object information, whereas the nonlinear image formed on the other side of the lens (517 nm) deviates in terms of size and location from the linear regime, obeying a modified Gaussian lens equation (*124*). Figure 5D shows scanning electron microscope (SEM) images of the fabricated lens in addition to an L-shaped aperture imaged using this lens. The longitudinal profiles of the fundamental and THG beams are shown in Fig. 5E. Notably, the THG image formed using this nonlinear scheme carries additional information about the spatial coherence of light emitted from the object, inferred from the high-order spatial correlations from different features of that object (not shown here).

More-complex behavior such as directional transmission based on asymmetric parametric conversion has been recently demonstrated. Devices of this kind can produce images in the visible spectral range when illuminated by infrared radiation while projecting different and independent images by reversing the direction of illumination (Fig. 5F) (*125*). Here, the resonators consist of two layers of materials: amorphous silicon and silicon nitride (Fig. 5G). Silicon has a higher refractive index and nonlinear susceptibility compared with silicon nitride. In this case, "forward" illumination leads to predominantly magnetic dipole–type scattering, whereas under "backward" illumination, the scattering is dominated by an electric dipole. The field enhancements within the nanoresonator are substantially higher for the forward excitation than they are for the

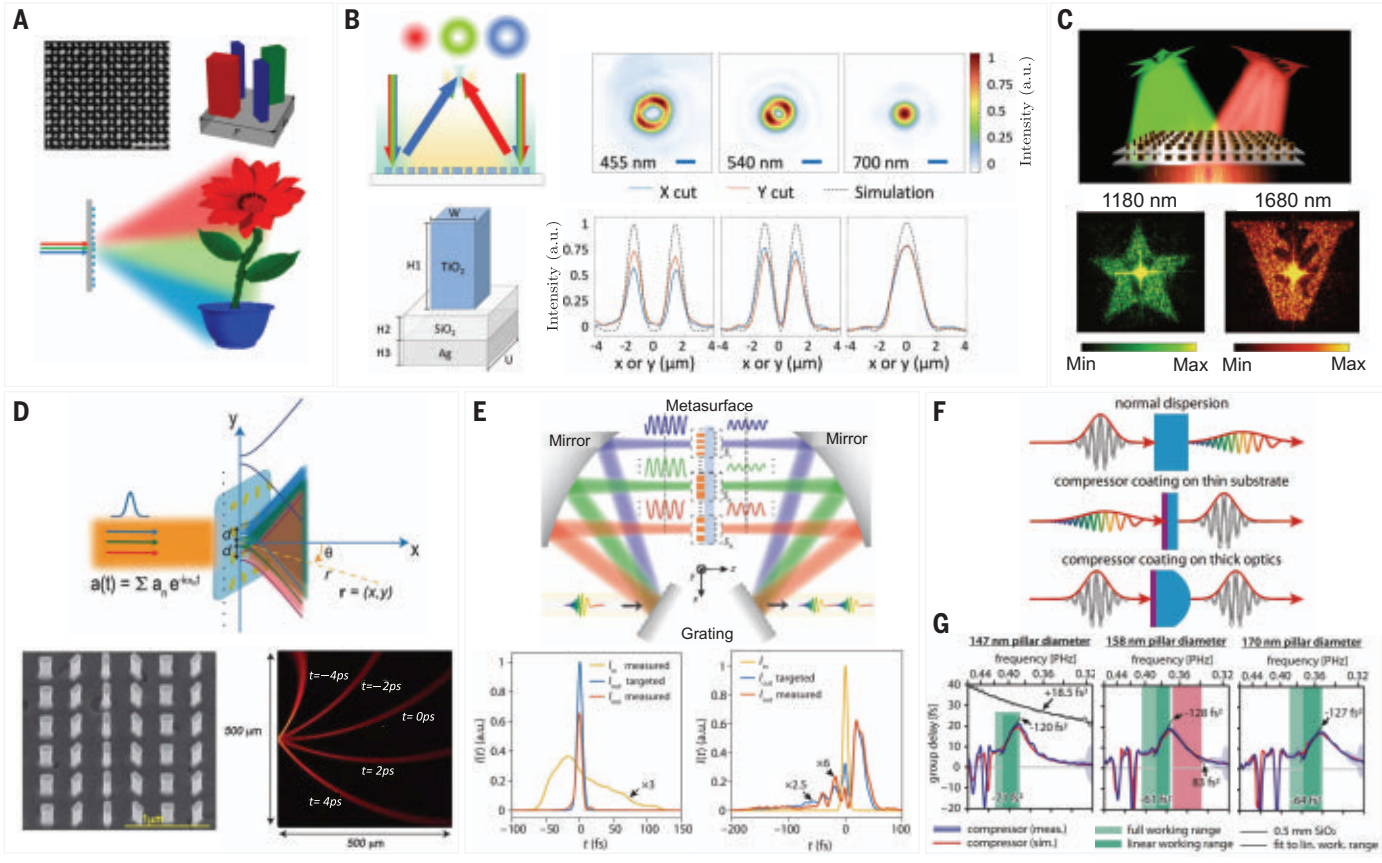


Fig. 4. Metasurface optics with multiwavelength control. (**A**) A highly dispersive meta-hologram can independently project distinct red, green, and blue images to reconstruct a holographic flower. The device relies on meta-molecules consisting of silicon nano blocks, each optimized for one wavelength, multiplexed altogether to form the metasurface. Partial SEM images of the fabricated metasurfaces are shown. Scale bar is 1 μm. *P*, periodicity of the metamolecule. Reprinted with permission from (*113*). Copyright 2016 American Chemical Society. (**B**) Wavelength-controlled beam generator focuses incoming red, green, and blue Gaussian beams into OAM states ℓ of 0, 1, and 2, respectively. The device relies on phase shifters made of titanium dioxide (TiO_2) square nanopillars on top of a silver (Ag) substrate with a thin layer of silicon dioxide (SiO_2) in between. The reflection phase can be controlled by adjusting the width of the nanopillar. Measured output intensity profiles are shown on the right (in 1D and 2D) for each incident wavelength. Scale bars are 2 μm. H, height; U, unit-cell length; W, width of the nanopillar. Reprinted with permission from (*115*). Copyright 2016 American Chemical Society. (**C**) Bilayer meta-holograms made of amorphous silicon nanopillars can project two independent (complex-amplitude modulated) logos in response to two different wavelengths, 1180 and 1680 nm. Images reproduced from (*116*). (**D**) Interaction between an input frequency-comb source and a passive metasurface where each spectral line is mapped to a spatial optical mode, via diffraction and focusing, generating a spatiotemporal optical pattern with time-varying tilt for beam-steering applications. a, incoming pulse composed of a discrete superposition of weighted spectral lines; d, spatial separation between the output spatial optical modes; r, position; t, time. Images reproduced from (*41*). (**E**) Ultrafast pulse shaping using metasurface-based Fourier transform. The setup consists of a pair of diffraction gratings, two parabolic mirrors, and a metasurface that is divided into *N* superpixels along the *x* direction to tailor the temporal characteristics of the output pulse. Here, *N* is 660 superpixels, each 34 μm in size. Temporal profiles of the targeted and measured output pulse are depicted at the bottom for a positively chirped input pulse (left) and a transform-limited input pulse (right). Images reproduced from (*117*). (**F**) Metasurface-based pulse compressor. Ultrashort pulses typically suffer from normal chromatic dispersion in transparent materials, which leads to pulse stretching. To mitigate these effects, a nanocoating (purple) can be applied to thin substrate to shorten elongated pulses or to thick optics to compensate for its group delay dispersion. Images reproduced from (*118*). (**G**) Measured and simulated compressor group delay profile compared with a fused silica substrate. The shaded regions mark spectral regions with different values of group delay dispersion. Images reproduced from (*118*).

backward direction. This leads to a drastic difference in brightness of the generated third-harmonic signal. Asymmetric harmonic generation can thus be used to generate two different images depending on the direction of illumination (Fig. 5H). This functionality can go beyond parametric generation of light and may find application in asymmetric generation of entangled photon states as well as realizing nonreciprocal behavior and optical isolation at the nanoscale.

Summary and outlook

Over the past decade, the use of flat optics has been widely extended from wavefront shaping and focusing to more sophisticated manipulation of structured light, owing to rich meta-atom libraries, accurate full-wave simulations, and precise nanofabrication. Complex configurations of single- and multilayer meta-atoms have enabled multifunctional behavior by simply changing one or more degrees of freedom of input light. This ability allows a static monolithically integrated photonic component to rapidly switch its behavior without the need for active circuitry. We have highlighted recent progress in this field as well as its possible applications in holography, structured-light generation, and polarization control. Metasurfaces with all-optical control knobs will serve as a key player in augmented reality and virtual reality (AR and VR) devices, 3D displays, and drone and automotive light detection and ranging (LIDAR) systems, owing to their light

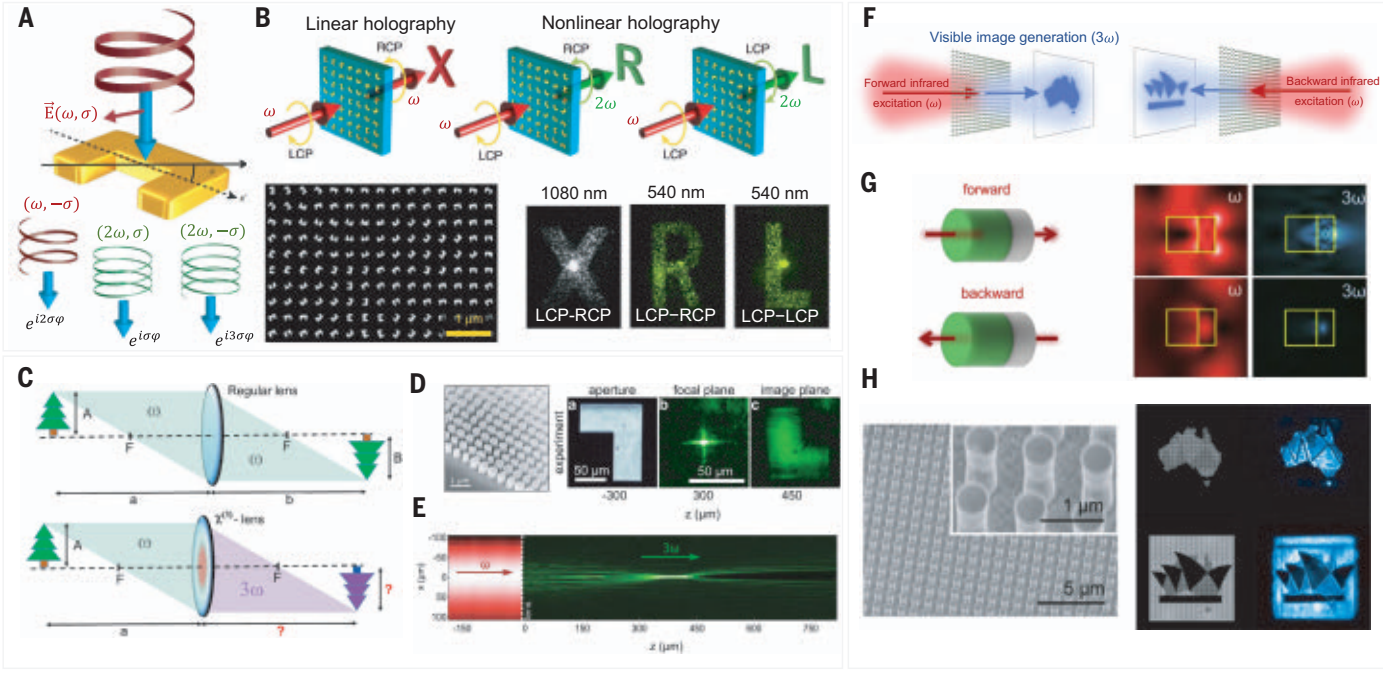


Fig. 5. Nonlinear metasurface holography and imaging. (**A**) Linear and nonlinear Berry phases from a gold split-ring resonator with an orientation angle of ϕ. While the transmitted fundamental cross–circularly polarized (CP) component accumulates a linear geometric phase of 2σϕ, CP and cross-CP components of the SHG incur nonlinear geometric phases of σϕ and 3σϕ, respectively. ω, angular frequency. Images reproduced from (*122*). (**B**) The linear channel of the meta-hologram projects the letter "X" at IR wavelengths, whereas the second-harmonic channels encode the letters "R" and "L" on different circular polarization states. SEM images of the device and its output response are shown. Images reproduced from (*122*). (**C**) Schematic of image formation with a regular lens compared with that with a nonlinear lens made of $\chi^{(3)}$ material, where the size and location of the formed image are modified. a, position of the object; A, height of the object; b, image position; B, image height; F, focal length. Reproduced with permission from (*124*). (**D**) An SEM image of the fabricated all-dielectric nonlinear metalens is shown on the left. A micrograph of the fabricated L-shaped aperture placed at an object distance −300 μm in front of the metalens is shown on the right. Although the conventional lens equation predicts image formation at infinity, the focal plane and the formed image of the THG lie at a *z* of 300 and 450 μm at the back focal plane of the metalens, respectively, governed by the generalized Gaussian lens equation. Reproduced with permission from (*124*). (**E**) Measured longitudinal intensity profiles of the fundamental (red) and THG (green) beams for the fundamental wavelength of 1550 nm. Reproduced with permission from (*124*). (**F**) Asymmetric parametric generation of images with a nonlinear metasurface. Different and independent THG images are generated in transmission, depending on the direction of illumination. Reproduced with permission from (*125*). (**G**) The metasurface is made of an anisotropic cylindrical nanoresonator consisting of silicon (gray) and silicon nitride (green) embedded into glass (left). Arrows visualize forward and backward excitation. Near-field distributions of the electric field at the fundamental (λ =1475 nm) and third-harmonic signal (λ = 492 nm) for the forward and backward directions show high transmission contrast (right). Yellow lines mark contours of the bilayer nanoresonator. Reproduced with permission from (*125*). (**H**) SEM images of the meta-hologram. Its nonlinear optical response detected in transmission at the third-harmonic frequency is shown on the right under forward and backward excitation at a wavelength of 1475 nm. Reproduced with permission from (*125*).

weight, ease of integration, and compact footprint. For instance, an autonomous vehicle can make use of metasurface-based polarization cameras to distinguish between different features on a road environment (e.g., water versus mud puddles) without the need for expensive and bulky camera systems (*89*). Additionally, nonlocal metalenses can be deployed in AR and VR wearable devices to reflect NIR light onto the iris for accurate eye tracking while transmitting visible light, ensuring unperturbed vision of the outside world (*126*). Moreover, integrating multifunctional metasurfaces with quantum emitters can enable efficient single-photon quantum sources and quantum-entangled photon states, whereas combining metasurfaces with single-photon-sensitive charge-coupled device cameras could allow multiple timeframe imaging and fast quantum measurements (*127*). Furthermore, complex states of classical and nonclassical structured light can be generated with metasurface-assisted laser cavities (*128*). The list goes on, with numerous applications that can make use of multifunctional flat optics, from laser beam shaping to optical communications, biomedical sensing, and imaging (*129*).

Nevertheless, several open challenges are still underway in metasurface research and applications. At the physical layer, topology-optimized meta-atom libraries seek to expand the function of forward-designed nanoantennas by exploring a larger design space of free-form geometries (*44*). Inverse-design and machine-learning techniques will be key tools for addressing this challenge by searching nonintuitive design spaces (*130*, *131*). More accurate models that account for the meta-atom coupling, in the absence of periodic boundary conditions, are also being tested now (*132*). Additionally, bilayer metasurfaces are being investigated as means for enhancing the diffraction efficiency (*133*) and realizing more versatile polarization control by relaxing the Jones matrix symmetry constraint imposed by their single-layer counterpart (*116*, *123*). With regard to mass production, several fabrication techniques are being developed to realize large area and conformal metasurfaces. From a light-wave standpoint, we expect that more attention will be given to the full vectorial nature of nonparaxial light (e.g., to control its longitudinal field component) as an additional degree of freedom (*134*). Likewise, optical coherence will be more heavily investigated as a fundamental property of light that acts as an additional control

knob if combined with engineering of the point spread function of the metasurface. In the near future, we envision that hybrid configurations of static and active metasurfaces will be integrated to realize the best of both worlds: high spatial resolution and time-varying response. With this versatile level of control, pushing structured light from 2D to 3D and from frozen to animate will help unlock a rich palette of optical phenomena throughout the next decade, from the atomic to the astrophysical scale.

REFERENCES AND NOTES

1. P. Vukusic, J. R. Sambles, Photonic structures in biology. *Nature* **424**, 852–855 (2003). doi: 10.1038/nature01941; pmid: 12917700
2. N. Capobianco *et al*., The *Grande* Rose of the Reims Cathedral: An eight-century perspective on the colour management of medieval stained glass. *Sci. Rep.* **9**, 3287 (2019). doi: 10.1038/s41598-019-39740-y; pmid: 30824744
3. L. Rayleigh, XVII. On the maintenance of vibrations by forces of double frequency, and on the propagation of waves through a medium endowed with a periodic structure. *London Edinb. Dublin Philos. Mag. J. Sci.* **24**, 145–159 (1887). doi: 10.1080/14786448708628074
4. E. Yablonovitch, Photonic band-gap structures. *J. Opt. Soc. Am. B* **10**, 283 (1993). doi: 10.1364/JOSAB.10.000283
5. E. Yablonovitch, Inhibited spontaneous emission in solid-state physics and electronics. *Phys. Rev. Lett.* **58**, 2059–2062 (1987). doi: 10.1103/PhysRevLett.58.2059; pmid: 10034639
6. S. John, Strong localization of photons in certain disordered dielectric superlattices. *Phys. Rev. Lett.* **58**, 2486–2489 (1987). doi: 10.1103/PhysRevLett.58.2486; pmid: 10034761
7. P. Lalanne, D. Lemercier-lalanne, On the effective medium theory of subwavelength periodic structures. *J. Mod. Opt.* **43**, 2063–2085 (1996). doi: 10.1080/09500349608232871
8. G. W. Milton, *Reformulating the Problem of Finding Effective Tensors* (Cambridge Univ. Press, 2002), pp. 245–270.
9. V. G. Veselago, The electrodynamics of substances with simultaneously negative values of ϵ and μ. *Sov. Phys. Usp.* **10**, 509–514 (1968). doi: 10.1070/PU1968v010n04ABEH003699
10. R. A. Shelby, D. R. Smith, S. Schultz, Experimental verification of a negative index of refraction. *Science* **292**, 77–79 (2001). doi: 10.1126/science.1058847; pmid: 11292865
11. J. Yao *et al*., Optical negative refraction in bulk metamaterials of nanowires. *Science* **321**, 930 (2008). doi: 10.1126/science.1157566; pmid: 18703734
12. R. W. Ziolkowski, N. Engheta, *Introduction, History, and Selected Topics in Fundamental Theories of Metamaterials* (Wiley, 2006), chap. 1, pp. 1–41.
13. M. Kadic, G. W. Milton, M. van Hecke, M. Wegener, 3D metamaterials. *Nat. Rev. Phys.* **1**, 198–210 (2019). doi: 10.1038/s42254-018-0018-y
14. A. M. Urbas *et al*., Roadmap on optical metamaterials. *J. Opt.* **18**, 093005 (2016). doi: 10.1088/2040-8978/18/9/093005
15. A. V. Kildishev, A. Boltasseva, V. M. Shalaev, Planar photonics with metasurfaces. *Science* **339**, 1232009 (2013). doi: 10.1126/science.1232009; pmid: 23493714
16. N. Yu, F. Capasso, Flat optics with designer metasurfaces. *Nat. Mater.* **13**, 139–150 (2014). doi: 10.1038/nmat3839; pmid: 24452357
17. D. Lin, P. Fan, E. Hasman, M. L. Brongersma, Dielectric gradient metasurface optical elements. *Science* **345**, 298–302 (2014). doi: 10.1126/science.1253213; pmid: 25035488
18. H.-T. Chen, A. J. Taylor, N. Yu, A review of metasurfaces: Physics and applications. *Rep. Prog. Phys.* **79**, 076401 (2016). doi: 10.1088/0034-4885/79/7/076401; pmid: 27308726
19. P. Genevet, F. Capasso, F. Aieta, M. Khorasaninejad, R. Devlin, Recent advances in planaroptics: From plasmonic to dielectric metasurfaces. *Optica* **4**, 139 (2017). doi: 10.1364/OPTICA.4.000139
20. S. M. Kamali, E. Arbabi, A. Arbabi, A. Faraon, A review of dielectric optical metasurfaces for wavefront control. *Nanophotonics* **7**, 1041–1068 (2018). doi: 10.1515/nanoph-2017-0129
21. T. W. Ebbesen, H. J. Lezec, H. F. Ghaemi, T. Thio, P. A. Wolff, Extraordinary optical transmission through sub-wavelength hole arrays. *Nature* **391**, 667–669 (1998). doi: 10.1038/35570
22. J. B. Pendry, Negative refraction makes a perfect lens. *Phys. Rev. Lett.* **85**, 3966–3969 (2000). doi: 10.1103/PhysRevLett.85.3966; pmid: 11041972
23. H. Liu, P. Lalanne, Microscopic theory of the extraordinary optical transmission. *Nature* **452**, 728–731 (2008). doi: 10.1038/nature06762; pmid: 18401405
24. N. Yu *et al*., Light propagation with phase discontinuities: Generalized laws of reflection and refraction. *Science* **334**, 333–337 (2011). doi: 10.1126/science.1210713; pmid: 21885733
25. P. Lalanne, S. Astilean, P. Chavel, E. Cambril, H. Launois, Design and fabrication of blazed binary diffractive elements with sampling periods smaller than the structural cutoff. *J. Opt. Soc. Am. A* **16**, 1143 (1999). doi: 10.1364/JOSAA.16.001143
26. Z. Bomzon, G. Biener, V. Kleiner, E. Hasman, Space-variant Pancharatnam-Berry phase optical elements with computer-generated subwavelength gratings. *Opt. Lett.* **27**, 1141–1143 (2002). doi: 10.1364/OL.27.001141; pmid: 18026387
27. M. Ozaki, J. Kato, S. Kawata, Surface-plasmon holography with white-light illumination. *Science* **332**, 218–220 (2011). doi: 10.1126/science.1201045; pmid: 21474756
28. P. Genevet, F. Capasso, Holographic optical metasurfaces: A review of current progress. *Rep. Prog. Phys.* **78**, 024401 (2015). doi: 10.1088/0034-4885/78/2/024401; pmid: 25609665
29. R. Zhao, L. Huang, Y. Wang, Recent advances in multi-dimensional metasurfaces holographic technologies. *PhotoniX* **1**, 20 (2020). doi: 10.1186/s43074-020-00020-y
30. M. Khorasaninejad *et al*., Metalenses at visible wavelengths: Diffraction-limited focusing and subwavelength resolution imaging. *Science* **352**, 1190–1194 (2016). doi: 10.1126/science.aaf6644; pmid: 27257251
31. M. Khorasaninejad, F. Capasso, Metalenses: Versatile multifunctional photonic components. *Science* **358**, eaam8100 (2017). doi: 10.1126/science.aam8100; pmid: 28982796
32. A. Arbabi, Y. Horie, M. Bagheri, A. Faraon, Dielectric metasurfaces for complete control of phase and polarization with subwavelength spatial resolution and high transmission. *Nat. Nanotechnol.* **10**, 937–943 (2015). doi: 10.1038/nnano.2015.186; pmid: 26322944
33. J. P. Balthasar Mueller, N. A. Rubin, R. C. Devlin, B. Groever, F. Capasso, Metasurface polarization optics: Independent phase control of arbitrary orthogonal states of polarization. *Phys. Rev. Lett.* **118**, 113901 (2017). doi: 10.1103/PhysRevLett.118.113901; pmid: 28368630
34. W. T. Chen, A. Y. Zhu, F. Capasso, Flat optics with dispersion-engineered metasurfaces. *Nat. Rev. Mater.* **5**, 604–620 (2020). doi: 10.1038/s41578-020-0203-3
35. Y. Hu *et al*., All-dielectric metasurfaces for polarization manipulation: Principles and emerging applications. *Nanophotonics* **9**, 3755–3780 (2020). doi: 10.1515/nanoph-2020-0220
36. N. Rubin, Z. Shi, F. Capasso, Polarization in diffractive optics and metasurfaces. *Adv. Opt. Photonics* **13**, 836 (2021). doi: 10.1364/AOP.439986
37. L. Huang, S. Zhang, T. Zentgraf, Metasurface holography: From fundamentals to applications. *Nanophotonics* **7**, 1169–1190 (2018). doi: 10.1515/nanoph-2017-0118
38. A. L. Holsteen, A. F. Cihan, M. L. Brongersma, Temporal color mixing and dynamic beam shaping with silicon metasurfaces. *Science* **365**, 257–260 (2019). doi: 10.1126/science.aax5961; pmid: 31320534
39. A. M. Shaltout, V. M. Shalaev, M. L. Brongersma, Spatiotemporal light control with active metasurfaces. *Science* **364**, eaat3100 (2019). doi: 10.1126/science.aat3100; pmid: 31097638
40. X. Guo, Y. Ding, Y. Duan, X. Ni, Nonreciprocal metasurface with space-time phase modulation. *Light Sci. Appl.* **8**, 123 (2019). doi: 10.1038/s41377-019-0225-z; pmid: 31871675
41. A. M. Shaltout *et al*., Spatiotemporal light control with frequency-gradient metasurfaces. *Science* **365**, 374–377 (2019). doi: 10.1126/science.aax2357; pmid: 31346064
42. W. Liu, Z. Li, H. Cheng, S. Chen, Dielectric resonance-based optical metasurfaces: From fundamentals to applications. *iScience* **23**, 101868 (2020). doi: 10.1016/j.isci.2020.101868; pmid: 33319185
43. K. Wu, P. Coquet, Q. J. Wang, P. Genevet, Modelling of free-form conformal metasurfaces. *Nat. Commun.* **9**, 3494 (2018). doi: 10.1038/s41467-018-05579-6; pmid: 30154424
44. Z. Shi *et al*., Continuous angle-tunable birefringence with freeform metasurfaces for arbitrary polarization conversion. *Sci. Adv.* **6**, eaba3367 (2020). doi: 10.1126/sciadv.aba3367; pmid: 32537506
45. A. Forbes, M. de Oliveira, M. R. Dennis, Structured light. *Nat. Photonics* **15**, 253–262 (2021). doi: 10.1038/s41566-021-00780-4
46. A. C. Overvig, S. C. Malek, N. Yu, Multifunctional nonlocal metasurfaces. *Phys. Rev. Lett.* **125**, 017402 (2020). doi: 10.1103/PhysRevLett.125.017402; pmid: 32678662
47. S. M. Kamali *et al*., Angle-multiplexed metasurfaces: Encoding independent wavefronts in a single metasurface under different illumination angles. *Phys. Rev. X* **7**, 041056 (2017). doi: 10.1103/PhysRevX.7.041056
48. X. Zhang *et al*., Ultrahigh-capacity dynamic holographic displays via anisotropic nanoholes. *Nanoscale* **9**, 1409–1415 (2017). doi: 10.1039/C6NR07854K; pmid: 28074963
49. S. Wan *et al*., Angular-multiplexing metasurface: Building up independent-encoded amplitude/phase dictionary for angular illumination. *Adv. Opt. Mater.* **9**, 2101547 (2021). doi: 10.1002/adom.202101547
50. X. Zhang *et al*., Controlling angular dispersions in optical metasurfaces. *Light Sci. Appl.* **9**, 76 (2020). doi: 10.1038/s41377-020-0313-0; pmid: 32411361
51. J. Jang, G.-Y. Lee, J. Sung, B. Lee, Independent multichannel wavefront modulation for angle multiplexed meta-holograms. *Adv. Opt. Mater.* **9**, 2100678 (2021). doi: 10.1002/adom.202100678
52. A. Leitis *et al*., Angle-multiplexed all-dielectric metasurfaces for broadband molecular fingerprint retrieval. *Sci. Adv.* **5**, eaaw2871 (2019). doi: 10.1126/sciadv.aaw2871; pmid: 31123705
53. E. Wang *et al*., Complete control of multichannel, angle-multiplexed, and arbitrary spatially varying polarization fields. *Adv. Opt. Mater.* **8**, 1901674 (2020). doi: 10.1002/adom.201901674
54. A. Silva *et al*., Performing mathematical operations with metamaterials. *Science* **343**, 160–163 (2014). doi: 10.1126/science.1242818; pmid: 24408430
55. C. Guo, M. Xiao, M. Minkov, Y. Shi, S. Fan, Photonic crystal slab laplace differentiator. *Optica* **5**, 251 (2017). doi: 10.1364/OPTICA.5.000251
56. H. Kwon, D. Sounas, A. Cordaro, A. Polman, A. Alù, Nonlocal metasurfaces for optical signal processing. *Phys. Rev. Lett.* **121**, 173004 (2018). doi: 10.1103/PhysRevLett.121.173004; pmid: 30411907
57. Y. Zhou, H. Zheng, I. I. Kravchenko, J. Valentine, Flat optics for image differentiation. *Nat. Photonics* **14**, 316–323 (2020). doi: 10.1038/s41566-020-0591-3
58. C. Guo, H. Wang, S. Fan, Squeeze free space with nonlocal flat optics. *Optica* **7**, 1133 (2020). doi: 10.1364/OPTICA.392978
59. O. Reshef *et al*., An optic to replace space and its application towards ultra-thin imaging systems. *Nat. Commun.* **12**, 3512 (2021). doi: 10.1038/s41467-021-23358-8; pmid: 34112771
60. C. Spägele *et al*., Multifunctional wide-angle optics and lasing based on supercell metasurfaces. *Nat. Commun.* **12**, 3787 (2021). doi: 10.1038/s41467-021-24071-2; pmid: 34145275
61. T. Kim *et al*., Asymmetric optical camouflage: Tuneable reflective colour accompanied by the optical Janus effect. *Light Sci. Appl.* **9**, 175 (2020). doi: 10.1038/s41377-020-00413-5; pmid: 33088492
62. Y. Chen, X. Yang, J. Gao, 3D Janus plasmonic helical nanoapertures for polarization-encrypted data storage. *Light Sci. Appl.* **8**, 45 (2019). doi: 10.1038/s41377-019-0156-8; pmid: 31098013
63. K. Chen *et al*., Directional Janus metasurface. *Adv. Mater.* **32**, e1906352 (2020). doi: 10.1002/adma.201906352; pmid: 31746042
64. G. Yoon, D. Lee, K. T. Nam, J. Rho, "Crypto-display" in dual-mode metasurfaces by simultaneous control of phase and spectral responses. *ACS Nano* **12**, 6421–6428 (2018). doi: 10.1021/acsnano.8b01344; pmid: 29924588
65. Z. Li *et al*., Full-space cloud of random points with a scrambling metasurface. *Light Sci. Appl.* **7**, 63 (2018). doi: 10.1038/s41377-018-0064-3; pmid: 30245810
66. E. Cohen *et al*., Geometric phase from Aharonov–Bohm to Pancharatnam–Berry and beyond. *Nat. Rev. Phys.* **1**, 437–449 (2019). doi: 10.1038/s42254-019-0071-1

67. Q. Hu, C. He, M. J. Booth, Arbitrary complex retarders using a sequence of spatial light modulators as the basis for adaptive polarisation compensation. *J. Opt.* **23**, 065602 (2021). doi: 10.1088/2040-8986/abed33
68. L. Nikolova, T. Todorov, Diffraction efficiency and selectivity of polarization holographic recording. *Opt. Acta* **31**, 579–588 (1984). doi: 10.1080/713821547
69. J. Tervo, J. Turunen, Paraxial-domain diffractive elements with 100% efficiency based on polarization gratings. *Opt. Lett.* **25**, 785–786 (2000). doi: 10.1364/OL.25.000785; pmid: 18064183
70. Z. Bomzon, G. Biener, V. Kleiner, E. Hasman, Radially and azimuthally polarized beams generated by space-variant dielectric subwavelength gratings. *Opt. Lett.* **27**, 285–287 (2002). doi: 10.1364/OL.27.000285; pmid: 18007778
71. Y. Zhao, A. Alu, Manipulating light polarization with ultrathin plasmonic metasurfaces. *Phys. Rev. B* **84**, 205428 (2011). doi: 10.1103/PhysRevB.84.205428
72. N. Yu *et al*., A broadband, background-free quarter-wave plate based on plasmonic metasurfaces. *Nano Lett.* **12**, 6328–6333 (2012). doi: 10.1021/nl303445u; pmid: 23130979
73. K. Y. Bliokh, F. J. Rodriguez-Fortuno, F. Nori, A. V. Zayats, Spin–orbit interactions of light. *Nat. Photonics* **9**, 796–808 (2015). doi: 10.1038/nphoton.2015.201
74. A. M. Yao, M. J. Padgett, Orbital angular momentum: Origins, behavior and applications. *Adv. Opt. Photonics* **3**, 161 (2011). doi: 10.1364/AOP.3.000161
75. L. Marrucci, C. Manzo, D. Paparo, Optical spin-to-orbital angular momentum conversion in inhomogeneous anisotropic media. *Phys. Rev. Lett.* **96**, 163905 (2006). doi: 10.1103/PhysRevLett.96.163905; pmid: 16712234
76. E. Maguid *et al*., Photonic spin-controlled multifunctional shared-aperture antenna array. *Science* **352**, 1202–1206 (2016). doi: 10.1126/science.aaf3417; pmid: 27103668
77. W.-T. Chen *et al*., High-efficiency broadband meta-hologram with polarization-controlled dual images. *Nano Lett.* **14**, 225–230 (2014). doi: 10.1021/nl403811d; pmid: 24329425
78. S. Kruk *et al*., Invited article: Broadband highly efficient dielectric meta-devices for polarization control. *APL Photonics* **1**, 030801 (2016). doi: 10.1063/1.4949007
79. S. Kruk *et al*., Transparent dielectric metasurfaces for spatial mode multiplexing. *Laser Photonics Rev.* **12**, 1800031 (2018). doi: 10.1002/lpor.201800031
80. E. Nazemosadat *et al*., Dielectric broadband metasurfaces for fiber mode multiplexed communications. *Adv. Opt. Mater.* **7**, 1801679 (2019). doi: 10.1002/adom.201801679
81. R. C. Devlin, A. Ambrosio, N. A. Rubin, J. P. B. Mueller, F. Capasso, Arbitrary spin-to-orbital angular momentum conversion of light. *Science* **358**, 896–901 (2017). doi: 10.1126/science.aao5392; pmid: 29097490
82. N. A. Rubin, A. Zaidi, A. H. Dorrah, Z. Shi, F. Capasso, Jones matrix holography with metasurfaces. *Sci. Adv.* **7**, eabg7488 (2021). doi: 10.1126/sciadv.abg7488; pmid: 34389537
83. M. Liu *et al*., Multifunctional metasurfaces enabled by simultaneous and independent control of phase and amplitude for orthogonal polarization states. *Light Sci. Appl.* **10**, 107 (2021). doi: 10.1038/s41377-021-00552-3; pmid: 34035215
84. Q. Fan *et al*., Independent amplitude control of arbitrary orthogonal states of polarization via dielectric metasurfaces. *Phys. Rev. Lett.* **125**, 267402 (2020). doi: 10.1103/PhysRevLett.125.267402; pmid: 33449781
85. Y. Yuan *et al*., Independent phase modulation for quadruplex polarization channels enabled by chirality-assisted geometric-phase metasurfaces. *Nat. Commun.* **11**, 4186 (2020). doi: 10.1038/s41467-020-17773-6; pmid: 32826879
86. R. Zhao *et al*., Multichannel vectorial holographic display and encryption. *Light Sci. Appl.* **7**, 95 (2018). doi: 10.1038/s41377-018-0091-0; pmid: 30510691
87. Y. Hu *et al*., Trichromatic and tripolarization-channel holography with noninterleaved dielectric metasurface. *Nano Lett.* **20**, 994–1002 (2020). doi: 10.1021/acs.nanolett.9b04107; pmid: 31880939
88. R. A. Chipman, W. S. T. Lam, G. Young, *Polarized Light and Optical Systems* (CRC Press, 2019).
89. N. A. Rubin *et al*., Matrix Fourier optics enables a compact full-Stokes polarization camera. *Science* **365**, eaax1839 (2019). doi: 10.1126/science.aax1839; pmid: 31273096
90. Z.-L. Deng *et al*., Diatomic metasurface for vectorial holography. *Nano Lett.* **18**, 2885–2892 (2018). doi: 10.1021/acs.nanolett.8b00047; pmid: 29590530
91. F. Ding *et al*., Versatile polarization generation and manipulation using dielectric metasurfaces. *Laser Photonics Rev.* **14**, 2000116 (2020). doi: 10.1002/lpor.202000116
92. E. Arbabi, S. M. Kamali, A. Arbabi, A. Faraon, Vectorial holograms with a dielectric metasurface: Ultimate polarization pattern generation. *ACS Photonics* **6**, 2712–2718 (2019). doi: 10.1021/acsphotonics.9b00678
93. Q. Song *et al*., Ptychography retrieval of fully polarized holograms from geometric-phase metasurfaces. *Nat. Commun.* **11**, 2651 (2020). doi: 10.1038/s41467-020-16437-9; pmid: 32461637
94. Z.-L. Deng *et al*., Full-color complex-amplitude vectorial holograms based on multi-freedom metasurfaces. *Adv. Funct. Mater.* **30**, 1910610 (2020). doi: 10.1002/adfm.201910610
95. D. Wen, J. J. Cadusch, J. Meng, K. B. Crozier, Vectorial holograms with spatially continuous polarization distributions. *Nano Lett.* **21**, 1735–1741 (2021). doi: 10.1021/acs.nanolett.0c04555; pmid: 33544611
96. A. Pors, M. G. Nielsen, S. I. Bozhevolnyi, Plasmonic metagratings for simultaneous determination of Stokes parameters. *Optica* **2**, 716 (2015). doi: 10.1364/OPTICA.2.000716
97. S. Wei, Z. Yang, M. Zhao, Design of ultracompact polarimeters based on dielectric metasurfaces. *Opt. Lett.* **42**, 1580–1583 (2017). doi: 10.1364/OL.42.001580; pmid: 28409803
98. E. Arbabi, S. M. Kamali, A. Arbabi, A. Faraon, Full-Stokes imaging polarimetry using dielectric metasurfaces. *ACS Photonics* **5**, 3132–3140 (2018). doi: 10.1021/acsphotonics.8b00362
99. A. H. Dorrah, N. A. Rubin, A. Zaidi, M. Tamagnone, F. Capasso, Metasurface optics for on-demand polarization transformations along the optical path. *Nat. Photonics* **15**, 287–296 (2021). doi: 10.1038/s41566-020-00750-2
100. A. H. Dorrah, N. A. Rubin, M. Tamagnone, A. Zaidi, F. Capasso, Structuring total angular momentum of light along the propagation direction with polarization-controlled meta-optics. *Nat. Commun.* **12**, 6249 (2021). doi: 10.1038/s41467-021-26253-4; pmid: 34716326
101. A. H. Dorrah, M. Zamboni-Rached, M. Mojahedi, Experimental demonstration of tunable refractometer based on orbital angular momentum of longitudinally structured light. *Light Sci. Appl.* **7**, 40 (2018). doi: 10.1038/s41377-018-0034-9; pmid: 30839632
102. Y. Qin, Y. Li, D. Deng, Y. Liu, M. Sun, Ultracompact biosensor based on a metalens with a longitudinally structured vector beam. *Appl. Opt.* **58**, 4438–4442 (2019). doi: 10.1364/AO.58.004438; pmid: 31251258
103. Q. Tian *et al*., The propagation properties of a longitudinal orbital angular momentum multiplexing system in atmospheric turbulence. *IEEE Photonics J.* **10**, 1–16 (2018). doi: 10.1109/JPHOT.2017.2778238
104. R. A. B. Suarez, L. A. Ambrosio, A. A. R. Neves, M. Zamboni-Rached, M. R. R. Gesualdi, Experimental optical trapping with frozen waves. *Opt. Lett.* **45**, 2514–2517 (2020). doi: 10.1364/OL.390909; pmid: 32356804
105. H. Ren *et al*., Metasurface orbital angular momentum holography. *Nat. Commun.* **10**, 2986 (2019). doi: 10.1038/s41467-019-11030-1; pmid: 31324755
106. H. Ren *et al*., Complex-amplitude metasurface-based orbital angular momentum holography in momentum space. *Nat. Nanotechnol.* **15**, 948–955 (2020). doi: 10.1038/s41565-020-0768-4; pmid: 32958936
107. X. Fang, H. Ren, M. Gu, Orbital angular momentum holography for high-security encryption. *Nat. Photonics* **14**, 102–108 (2020). doi: 10.1038/s41566-019-0560-x
108. H. Zhou *et al*., Polarization-encrypted orbital angular momentum multiplexed metasurface holography. *ACS Nano* **14**, 5553–5559 (2020). doi: 10.1021/acsnano.9b09814; pmid: 32348122
109. G. Qu *et al*., Reprogrammable meta-hologram for optical encryption. *Nat. Commun.* **11**, 5484 (2020). doi: 10.1038/s41467-020-19312-9; pmid: 33127918
110. P. Georgi *et al*., Optical secret sharing with cascaded metasurface holography. *Sci. Adv.* **7**, eabf9718 (2021). doi: 10.1126/sciadv.abf9718; pmid: 33853788
111. S. W. D. Lim, J.-S. Park, M. L. Meretska, A. H. Dorrah, F. Capasso, Engineering phase and polarization singularity sheets. *Nat. Commun.* **12**, 4190 (2021). doi: 10.1038/s41467-021-24493-y; pmid: 34234140
112. G. H. Yuan, N. I. Zheludev, Detecting nanometric displacements with optical ruler metrology. *Science* **364**, 771–775 (2019). doi: 10.1126/science.aaw7840; pmid: 31072905
113. B. Wang *et al*., Visible-frequency dielectric metasurfaces for multiwavelength achromatic and highly dispersive holograms. *Nano Lett.* **16**, 5235–5240 (2016). doi: 10.1021/acs.nanolett.6b02326; pmid: 27398793
114. O. Avayu, E. Almeida, Y. Prior, T. Ellenbogen, Composite functional metasurfaces for multispectral achromatic optics. *Nat. Commun.* **8**, 14992 (2017). doi: 10.1038/ncomms14992; pmid: 28378810
115. Z. Shi *et al*., Single-layer metasurface with controllable multiwavelength functions. *Nano Lett.* **18**, 2420–2427 (2018). doi: 10.1021/acs.nanolett.7b05458; pmid: 29461838
116. Y. Zhou *et al*., Multifunctional metaoptics based on bilayer metasurfaces. *Light Sci. Appl.* **8**, 80 (2019). doi: 10.1038/s41377-019-0193-3; pmid: 31666946
117. S. Divitt, W. Zhu, C. Zhang, H. J. Lezec, A. Agrawal, Ultrafast optical pulse shaping using dielectric metasurfaces. *Science* **364**, 890–894 (2019). doi: 10.1126/science.aav9632; pmid: 31048550
118. M. Ossiander *et al*., Slow light nanocoatings for ultrashort pulse compression. *Nat. Commun.* **12**, 6518 (2021). doi: 10.1038/s41467-021-26920-6; pmid: 34764297
119. G. Li, S. Zhang, T. Zentgraf, Nonlinear photonic metasurfaces. *Nat. Rev. Mater.* **2**, 17010 (2017). doi: 10.1038/natrevmats.2017.10
120. T. Pertsch, Y. Kivshar, Nonlinear optics with resonant metasurfaces. *MRS Bull.* **45**, 210–220 (2020). doi: 10.1557/mrs.2020.65
121. A. Krasnok, M. Tymchenko, A. Alu, Nonlinear metasurfaces: A paradigm shift in nonlinear optics. *Mater. Today* **21**, 8–21 (2018). doi: 10.1016/j.mattod.2017.06.007
122. W. Ye *et al*., Spin and wavelength multiplexed nonlinear metasurface holography. *Nat. Commun.* **7**, 11930 (2016). doi: 10.1038/ncomms11930; pmid: 27306147
123. E. Almeida, O. Bitton, Y. Prior, Nonlinear metamaterials for holography. *Nat. Commun.* **7**, 12533 (2016). doi: 10.1038/ncomms12533; pmid: 27545581
124. C. Schlickriede *et al*., Nonlinear imaging with all-dielectric metasurfaces. *Nano Lett.* **20**, 4370–4376 (2020). doi: 10.1021/acs.nanolett.0c01105; pmid: 32374616
125. S. Kruk *et al*., *Asymmetric parametric generation of images with nonlinear dielectric metasurfaces. arXiv:2108.04425 [physics.optics] (2021).*
126. J.-H. Song, J. van de Groep, S. J. Kim, M. L. Brongersma, Non-local metasurfaces for spectrally decoupled wavefront manipulation and eye tracking. *Nat. Nanotechnol.* **16**, 1224–1230 (2021). doi: 10.1038/s41565-021-00967-4; pmid: 34594006
127. A. S. Solntsev, G. S. Agarwal, Y. S. Kivshar, Metasurfaces for quantum photonics. *Nat. Photonics* **15**, 327–336 (2021). doi: 10.1038/s41566-021-00793-z
128. H. Sroor *et al*., High-purity orbital angular momentum states from a visible metasurface laser. *Nat. Photonics* **14**, 498–503 (2020). doi: 10.1038/s41566-020-0623-z
129. C.-W. Qiu, T. Zhang, G. Hu, Y. Kivshar, Quo vadis, metasurfaces? *Nano Lett.* **21**, 5461–5474 (2021). doi: 10.1021/acs.nanolett.1c00828; pmid: 34157842
130. C. M. Lalau-Keraly, S. Bhargava, O. D. Miller, E. Yablonovitch, Adjoint shape optimization applied to electromagnetic design. *Opt. Express* **21**, 21693–21701 (2013). doi: 10.1364/OE.21.021693; pmid: 24104043
131. L. Li *et al*., Machine-learning reprogrammable metasurface imager. *Nat. Commun.* **10**, 1082 (2019). doi: 10.1038/s41467-019-09103-2; pmid: 30842417
132. M. Mansouree, A. McClung, S. Samudrala, A. Arbabi, Large-scale parametrized metasurface design using adjoint optimization. *ACS Photonics* **8**, 455–463 (2021). doi: 10.1021/acsphotonics.0c01058
133. M. Mansouree *et al*., Multifunctional 2.5D metastructures enabled by adjoint optimization. *Optica* **7**, 77 (2020). doi: 10.1364/OPTICA.374787
134. H. Ren, W. Shao, Y. Li, F. Salim, M. Gu, Three-dimensional vectorial holography based on machine learning inverse design. *Sci. Adv.* **6**, eaaz4261 (2020). doi: 10.1126/sciadv.aaz4261; pmid: 32494614

ACKNOWLEDGMENTS

Funding: We acknowledge financial support from the National Science Foundation (grant no. ECCS-2025158), Office of Naval Research (grant no. N00014-20-1-2450), Air Force Office of Scientific Research (grant no. FA95550-19-1-0135), and Natural Sciences and Engineering Research Council of Canada (grant no. PDF-533013-2019). **Competing interests:** The authors declare that they have no competing interests.

10.1126/science.abi6860

AAAS.ORG/COMMUNITY

AAAS' Member Community is a one-stop destination for scientists and STEM enthusiasts alike. It's "Where Science Gets Social": a community where facts matter, ideas are big and there's always a reason to come hang out, share, discuss and explore.

Member COMMUNITY
AAAS

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

RESEARCH ARTICLE SUMMARY

CANCER GENOMICS

Substitution mutational signatures in whole-genome–sequenced cancers in the UK population

Andrea Degasperi, Xueqing Zou, Tauanne Dias Amarante, Andrea Martinez-Martinez, Gene Ching Chiek Koh, João M. L. Dias, Laura Heskin, Lucia Chmelova, Giuseppe Rinaldi, Valerie Ya Wen Wang, Arjun S. Nanda, Aaron Bernstein, Sophie E. Momen, Jamie Young, Daniel Perez-Gil, Yasin Memari, Cherif Badja, Scott Shooter, Jan Czarnecki, Matthew A. Brown, Helen R. Davies, Genomics England Research Consortium, Serena Nik-Zainal*

INTRODUCTION: Mutational signatures—imprints of DNA damage and repair processes that have been operative during tumorigenesis—provide insights into environmental and endogenous causes of each patient's cancer. Cancer genome sequencing studies permit exploration of mutational signatures. We investigated a very large number of whole-genome–sequenced cancers of many tumor types, substantially more than in previous efforts, to comprehensively reinforce our understanding of mutational signatures.

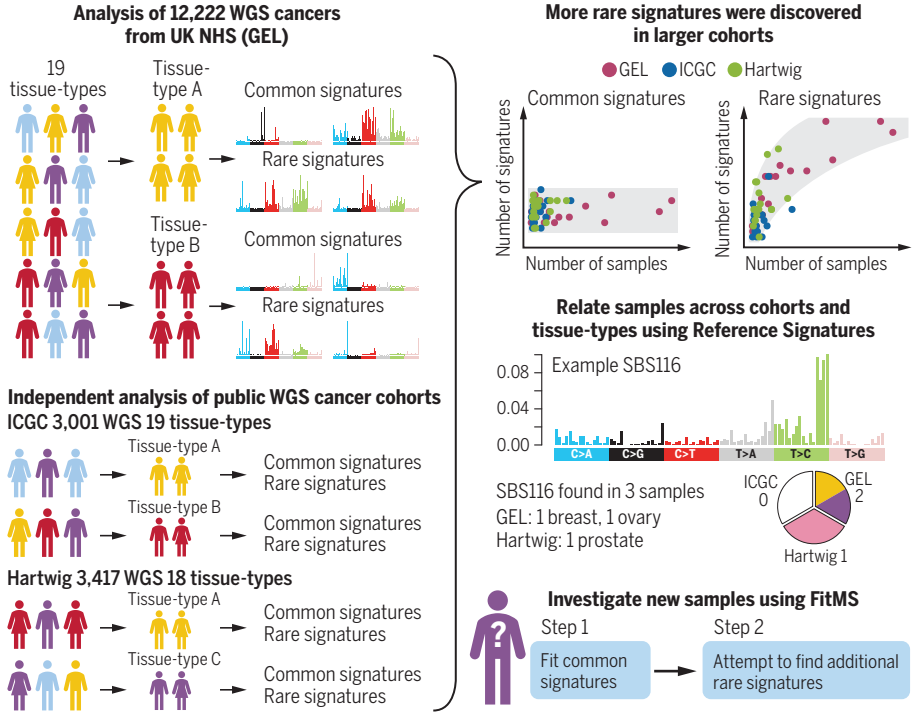
RATIONALE: We present mutational signature analyses of 12,222 whole-genome–sequenced cancers collected prospectively via the UK National Health Service (NHS) for the 100,000 Genomes Project. We identified single-base substitution (SBS) and double-base substitution (DBS) signatures independently in each organ. Exploiting this unusually large cohort, we developed a method to enhance discrimination of common mutational processes from rare, lower-frequency mutagenic processes. We validated our findings by independently performing analyses with data from two publicly available cohorts: 3001 primary cancers from the International Cancer Genome Consortium (ICGC) and 3417 metastatic cancers from the Hartwig Medical Foundation. We produced a set of reference signatures by comparing and contrasting the independently derived tissue-specific signatures and performing clustering analysis to unite mutational signatures from different tissues that could be due to similar processes. We included additional quality control measures such as dimensionality reduction of mixed signatures and gathered evidence that could help elucidate mechanisms and etiologies such as transcriptional and replication strand bias, associations with somatic drivers, and germline predisposition mutations. We also investigated additional mutation context and examined past clinical and treatment histories when possible, to explore potential etiologies.

RESULTS: Each organ contained a limited number of common SBS signatures (typically between 5 and 10). The number of common signatures was independent of cohort size. By contrast, the number of rare signatures was dependent on sample size, as the likelihood of detecting a rare signature is a function of its population prevalence. The same biological process produced slightly different signatures in diverse tissues, reinforcing that mutational signatures are tissue specific.

Across organs, we clustered all tissue-specific signatures to ascertain mutational processes that were equivalent but occurring in different tissues (i.e., reference signatures). We obtained 82 high-confidence SBS reference signatures and 27 high-confidence DBS reference signatures. We compared these with previously reported mutational signatures, revealing 40 and 18 previously unidentified SBS and DBS signatures, respectively.

Because we are cognizant of increasing complexity in mutational signatures and want to enable general users, we developed an algorithm called Signature Fit Multi-Step (FitMS) that seeks signatures in new samples while taking advantage of our recent findings. In a first step, FitMS detects common, organ-specific signatures; in a second step, it determines whether an additional rare signature is also present.

CONCLUSION: Mutational signature analysis of 18,640 cancers, the largest cohort of whole-genome–sequenced samples to date, has required methodological advances, permitting knowledge expansion. We have identified many previously unreported signatures and established the concept of common and rare signatures. The FitMS algorithm has been designed to exploit these advances to aid users in accurately identifying mutational processes in new samples. ■



Discovery and application of common and rare mutational signatures. Analysis of three large whole-genome–sequenced cancer cohorts revealed that per-organ common signatures are limited in number, whereas numbers of rare signatures increase with increasing cohort size. Reference signatures permit comparisons across organs and cohorts. Henceforth, a new algorithm, FitMS, which accounts for common and rare signatures, can be used to analyze new samples. GEL, Genomics England cohort.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: sn206@cam.ac.uk
Cite this article as A. Degasperi *et al.*, *Science* **376**, eabl9283 (2022). DOI: 10.1126/science.abl9283

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abl9283

RESEARCH ARTICLE

CANCER GENOMICS

Substitution mutational signatures in whole-genome–sequenced cancers in the UK population

Andrea Degasperi[1,2], Xueqing Zou[1,2], Tauanne Dias Amarante[1,2], Andrea Martinez-Martinez[1,2], Gene Ching Chiek Koh[1,2], João M. L. Dias[1,2], Laura Heskin[1,2], Lucia Chmelova[1,2], Giuseppe Rinaldi[1,2], Valerie Ya Wen Wang[1,2], Arjun S. Nanda[1,2], Aaron Bernstein[1,2], Sophie E. Momen[1,2], Jamie Young[1,2], Daniel Perez-Gil[1,2], Yasin Memari[1,2], Cherif Badja[1,2], Scott Shooter[1,2], Jan Czarnecki[1,2], Matthew A. Brown[3,4], Helen R. Davies[1,2], Genomics England Research Consortium[3]†, Serena Nik-Zainal[1,2]*

Whole-genome sequencing (WGS) permits comprehensive cancer genome analyses, revealing mutational signatures, imprints of DNA damage, and repair processes that have arisen in each patient's cancer. We performed mutational signature analyses on 12,222 whole-genome–sequenced tumor-normal matched pairs from patients recruited via the UK National Health Service (NHS). We contrasted our results with two independent cancer WGS datasets—from the International Cancer Genome Consortium (ICGC) and the Hartwig Medical Foundation (HMF)—involving 18,640 whole-genome–sequenced cancers in total. Our analyses add 40 single and 18 double substitution signatures to the current mutational signature tally. We show for each organ that cancers have a limited number of common signatures and a long tail of rare signatures, and we provide a practical solution for applying this concept of common versus rare signatures to future analyses.

The global cancer burden was estimated at 19.3 million new cases and 10.0 million deaths in 2020 (*1*). Worldwide, cancer is the first or second leading cause of mortality before the age of 70 (*1*). The genome of a cancer is a highly distorted entity that has acquired thousands of genetic aberrations since conception (*2*). If examined comprehensively, cancer genomes can thus reveal insights into carcinogenesis (*2*).

Today, modern sequencing technologies have augmented the scale and rapidity of genome resequencing (*3*), permitting whole-genome sequencing (WGS) approaches that provide an all-inclusive perspective on cancer genomes (*4*). Beyond the handful of causative driver mutations, WGS allows exploration of the full landscape of passenger mutations that describe the processes that have arisen during tumorigenesis, resulting in patterns known as mutational signatures (*5*–*7*). Whereas drivers become important targets for therapeutic intervention, mutational signatures provide clues regarding historical environmental exposures and highlight potentially targetable pathway defects (*4*, *6*, *8*, *9*).

Substantial efforts by The Cancer Genome Atlas (TCGA) (*10*), the International Cancer Genome Consortium (ICGC) (*9*, *11*), and the Hartwig Medical Foundation (HMF) (*12*) have helped advance cancer genomics considerably in recent years. However, an endeavor to generate whole cancer genomes from national public health cancer services would be a welcome demonstration of how cancer genomic data can be derived in real time and ultimately benefit patients and the scientific community.

We examined a cohort of 15,838 whole-genome–sequenced cancers from patients recruited from all 13 UK National Health Service (NHS) Genomic Medicine Centres across England as part of the Genomics England (GEL) 100,000 Genomes Project (100kGP) (*7*, *13*) (GEL v8 data release). We report the analysis of mutational signatures and highlight a conceptual advance achieved via the examination of this substantial WGS collection. We add 40 single-base substitution (SBS) mutational signatures and 18 double-base substitution (DBS) mutational signatures to the current tally. We compare these additional signatures to known etiologies and end by suggesting principles of how to meaningfully use mutational signatures in future analyses.

The GEL cohort

All 15,838 tumor-normal sample pairs were taken through 100kGP bioinformatic somatic-variant analysis pipelines. We restricted our analysis to high-quality data derived from flash-frozen material, involving 12,222 GEL tumor samples from 11,585 individuals (several participants had synchronous or metachronous tumors). For this evaluation, the final dataset included a total of 298,694,545 substitutions, 2,675,617 double substitutions, 154,675,475 indels, and 1,958,105 rearrangements (Fig. 1, A and B, and tables S1 and S2) of 19 tumor types [skin, lung, stomach, colorectal, bladder, liver, uterus, ovary, biliary, kidney, pancreas, breast, prostate, bone and soft tissue, central nervous system (CNS), lymphoid, oropharyngeal, neuroendocrine tumors (NETs), and myeloid].

Common and rare mutational signatures

The national GEL sequencing endeavor delivers thousands of samples for certain tumor types (1009 lung, 1355 kidney, 2572 breast, and 1480 bone and soft-tissue cancers), one or two orders of magnitude greater than the number of samples generated from previous WGS efforts for some organs. This permits robust detection of signatures that are rare—i.e., those occurring in ≤1% of the tumors. Furthermore, already-sequenced WGS cohorts, such as ~3000 primary cancers from ICGC and ~3400 metastatic cancers from HMF, provide a powerful means of validating findings.

We performed mutational signature extractions confined to specific tumor types by means of an updated signature extraction method (Fig. 1C, fig. S1, tables S3 to S6, and materials and methods). Briefly, for each tumor type, we clustered mutational catalogs (counts of SBSs in 96-element form or DBSs in 78-element form), selecting only samples with recurrent, commonly occurring profiles to perform signature extraction (fig. S1, A to C). Cases with unusual profiles and likely to have rare signatures were excluded in the first extraction, thus yielding a set of highly accurate common signatures that are prevalent for a given tumor type. Next, by fitting these common signatures into all samples, cases that are likely to have additional patterns not fully explained by common signatures alone would report a high error (or discrepancy between true sample catalog and reconstructed catalog) (fig. S1D). Potential additional signatures were then extracted from these samples to obtain a set of rare signatures (fig. S1, E to H, and materials and methods). Accordingly, we obtained a set of common signatures and a set of rare signatures for each tumor type. In all, for SBSs, we identified 135 common signatures and 180 rare signatures in 19 tumor types within the GEL cancer cohort.

To validate these common and rare signatures, we performed signature extractions in independent cohorts of 3001 ICGC primary whole-genome–sequenced cancers (19 tumor types) and 3417 metastatic HMF WGS samples (18 tumor types). We identified 135 common and 58 rare signatures in the ICGC cohort and 135 common and 114 rare signatures in the HMF cohort (tables S7 to S10). We performed an agnostic three-way signature comparison in 16 tissue types that were present in all three cohorts (fig. S2) and found that signatures

[1]Academic Department of Medical Genetics, School of Clinical Medicine, University of Cambridge, Cambridge CB2 0QQ, UK. [2]Early Cancer Institute, University of Cambridge, Cambridge CB2 0XZ, UK. [3]Genomics England, Queen Mary University of London, Dawson Hall, Charterhouse Square, London EC1M 6BQ, UK. [4]Faculty of Life Sciences and Medicine, King's College London, London SE1 9RT, UK.
*Corresponding author. Email: sn206@cam.ac.uk
†Genomics England Research Consortium members and affiliations are listed at the end of this paper.

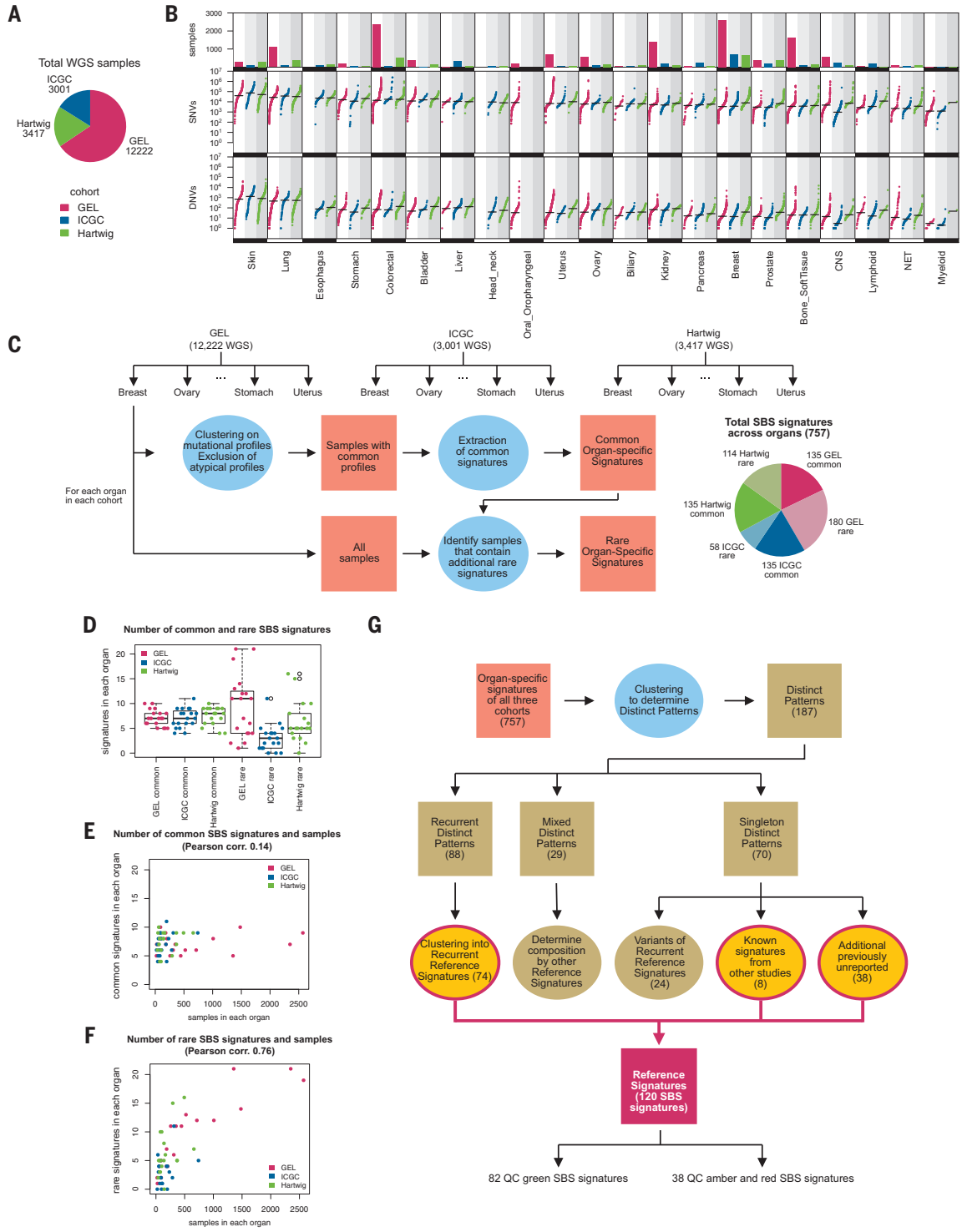


Fig. 1. Whole-genome–sequenced cancers across three independent cohorts: GEL, ICGC, and HMF. (**A**) WGS cases included in analyses. (**B**) Number of samples and mutational burden of somatic single-nucleotide variants (SNVs) and double-nucleotide variants (DNVs) across 21 whole-genome–sequenced tumor types from the GEL, ICGC, and HMF cohorts. Not all tumor types (e.g., esophagus, head and neck, oropharyngeal) are represented in all three cohorts. (**C**) Schematic representation of the workflow of mutational signature analysis. Three cohorts (GEL, ICGC, and HMF) were evaluated independently. For each organ in each cohort, mutational catalogs were clustered and samples with atypical catalogs were excluded from the extraction process. Samples with similar catalogs were subjected to signature extraction to obtain a set of common organ-specific signatures. These common signatures were fitted into all samples, highlighting samples with a high error profile that were subsequently used to identify rare signatures. The pie chart shows the total number of SBS signatures identified for each independent extraction of each organ in all three cohorts. (**D**) Number of common and rare SBS signatures in each cohort. (**E**) Common SBS signatures as a function of number of samples analyzed. (**F**) Rare SBS signatures as a function of number of samples analyzed. (**G**) Procedure to determine the reference signatures from all identified cohort-organ signatures. Numbers refer to the SBS signature analysis. For details, see materials and methods.

from the same organ in different cohorts were more similar to each other than to those in other tissue types, providing reassuring evidence that mutational signatures in each organ are highly reproducible, have tissue specificities, and were detectable regardless of sequencing platform or mutation-calling algorithm (fig. S2).

The number of common signatures in each organ is usually limited (between 5 and 10 for SBSs) and is independent of the number of samples analyzed per organ (Fig. 1, D and E; fig. S3, A and B; and tables S11 and S12). By contrast, the number of rare signatures varies and is highly correlated with the number of samples analyzed (Fig. 1F and fig. S3C). This illuminates why ubiquitous, organ-specific signatures are detectable even with limited numbers of whole genomes, whereas sporadic, rare signatures are more likely to be detected with increased sample size.

Reference mutational signatures

Biologically, the same mutational processes could underpin signatures extracted from different tumor types. Thus, we considered all common and rare GEL, ICGC, and HMF tissue-specific signatures together, involving 18,640 whole-genome–sequenced cancer samples (Fig. 1G and materials and methods), and performed a clustering analysis to derive a set of reference signatures. First, we identified clusters of highly similar patterns that we termed "distinct patterns" (tables S13 to S16). Each distinct pattern could be (i) a true signature, thus observable in independent extractions of diverse organs and cohorts (recurrent pattern); (ii) a mix of other signatures (mixed pattern); or (iii) a pattern seen in only one extraction (singleton pattern) (Fig. 1G and tables S17 and S18). Next, we determined a minimal set of reference signatures, which were classified as quality control (QC) green, amber, or red (green implies high-confidence signatures observed in multiple independent extractions; amber and red signatures were observed only once or were possible mathematical artifacts) (Fig. 1G and tables S19 to S22). In all, we identified 82 SBS and 27 DBS high-quality signatures (figs. S4 to S6). Henceforth, we will discuss only the high-quality QC green signatures, although all signatures are available for reference in the supplementary materials (tables S19 and S20).

Reference signatures were compared and matched with Catalogue of Somatic Mutations in Cancer (COSMIC) mutational signatures (*14*), confirming 42 and 9 previously described COSMIC SBS and DBS signatures, respectively (Fig. 2, A and B; figs. S3, D to G, S4, and S6A; and table S19). We found 40 previously unreported high-confidence SBS signatures and 18 previously unreported DBS signatures in this analysis (figs. S5 and S6B). Consistent with prior nomenclature (*14*), these SBS signatures have been numbered from 95 onward and the DBS signatures from 12 onward (table S19). Note that COSMIC and/or reference signatures are a simplified means of discussing signatures that are mutually present across tissues. However, they are purely mathematical constructs—an averaged result across different organs—and thus organ-specific signatures are more likely to be accurate biological representations of the mutational processes that occur within a tissue. We also provide the numbers of mutations associated with each reference signature per sample (tables S23 and S24) and matrices to map each reference signature to organ-specific signatures (tables S25 and S26) for SBSs and DBSs, respectively.

Previously unreported mutational signatures

Single-base substitution signatures

We note four previously unreported and five recently reported (*15*–*17*) signatures that are common, recurring in many samples of multiple tumor types in all three cohorts (GEL, ICGC, and HMF), and detectable because of the scale of this analysis (Fig. 2C). Among the previously unreported signatures, SBS107 is dominated by C>A variants and reported consistently in kidney and bladder cancers, suggestive of an organ-specific process. SBS100 bears similarities to the APOBEC signature SBS2; however, it presents a taller TCC>TTC peak (underlines indicate mutated bases) and additional context-independent C>T mutations. SBS110 has the tallest T>A peak at CTG>CAG, with contributions from T>C at ATA and ATG. The preponderance in the liver and biliary tract would suggest a compound that is likely cleared through the hepatobiliary system. SBS121 is characterized by C>G variants mostly at ACT and TCT contexts, shows replication strand bias (RSB), and is found mostly in colorectal and stomach cancers. We also confirm the recently reported SBS92 (*15*), SBS93, SBS94 (*16*), SBS125, and SBS127 [RefSig N12 and N1, respectively (*17*)].

Three signatures occurred frequently in specific tumor types (Fig. 2D): SBS120 dominated by T>C mutations at ATN and a distinctive peak of C>T at GCG, seen in 75% of CNS cancers; SBS122 characterized by T>C mutations in general but primarily TTN, in 67% of sarcomas; and SBS101 defined by C>T variants, in 68% of lymphoid cancers.

Thirty-one additional rare, previously unreported signatures of high confidence were present in ≤1% of samples (Fig. 2E). We discuss several in detail in relevant sections below and, for brevity, tabulate most of them in table S19 (associated information, such as transcriptional and replication strand asymmetries, is included there). All mutational signature data can also be viewed at our website, Signal: https://signal.mutationalsignatures.com/explore/study/6.

Double- and triple-base substitution signatures

We adopted similar principles to identify 39 DBSs, including 27 of high confidence (materials and methods, table S20, and fig. S6). We performed three additional evaluations: First, we curated dinucleotides for each DBS signature in the GEL dataset to check that they were in cis (i.e., on the same DNA strand). Second, for a DBS signature that was correlated with an SBS signature, an in silico analysis was performed to assess whether the DBS pattern could be expected given the SBS pattern (materials and methods). Third, we investigated up to 10 nucleotides (nt) of mutational context of relevant dinucleotides for each DBS signature. These assessments were critical in refuting several DBS signatures as being simply due to chance, as described below.

Of 11 previously described COSMIC DBS signatures (*14*), we identified 9 and were unable to extract DBS6 or DBS9 (Fig. 3A and figs. S3, F and G, and S6A). Of our 27 high-confidence DBS signatures, 17 had bona fide dinucleotides in cis. We confirmed previously reported signatures and their associated etiologies: DBS1 [ultraviolet (UV) light], DBS2 (smoking), DBS5 (platinum therapy), and DBS11 (associated with APOBEC; here verified as APOBEC-induced given the 10-nt sequence context analysis showing a TpCC preponderance) (Fig. 3B). DBS7 was previously reported as associated with MMR defects (*14*), whereas we find associations with SBS17 instead (fig. S7A). DBS8, mostly in colorectal cancer, showed dinucleotide variants often preceded by cytosine and followed by adenine (fig. S7, B and C).

We confirm that DBS5 and DBS18 are associated with prior platinum exposure (*18*). Mutational context analysis indicates that these are distinct signatures: DBS5 has the tallest peak of CT>AA mutations without preference in flanking sequences, whereas DBS18 has the tallest peak of CT>AC mutations, in which the dinucleotide is always preceded by cytosine (fig. S7D). Both signatures have a TG>GT peak most frequently followed by guanine (fig. S7, E and F).

DBS13 and DBS20 were low-burden signatures that appear to correlate with each other and SBS8 (Fig. 3, C and D). DBS16 was associated with SBS10d (Fig. 3C), a hypermutator signature recently reported as due to polymerase δ (POLD) dysfunction (*19*). DBS22 is not associated with very prominent peaks (highest probabilities only 7%). However, it appears to be correlated with SBS9 and is seen only in lymphoid cancers (Fig. 3, C and D). DBS26 is similar to DBS7 and correlates with SBS17 in esophageal and stomach cancers (Fig. 3, C and D). DBS30 was observed in one lymphoid cancer sample and may be related to treatment (fig. S6B).

DBS25 is characterized by an excess of TT>AA that, on inspection, reveals a triple-base

Fig. 2. SBS signatures across 18,640 whole-genome–sequenced cancers. (**A**) Frequency of SBS signatures in the present study. Orange bars highlight 42 signatures reported in this study and present in COSMIC v3.2; blue bars highlight 40 previously unreported signatures found in the present analysis. (**B**) Same information as in (A), with a log scale on the *y* axis. (**C**) Four previously unreported (SBS107, SBS100, SBS110, and SBS121) and five recently reported (SBS92, SBS93, SBS94, SBS125, and SBS127) common SBS signatures found in many organ systems. (**D**) Previously unreported common SBS signatures found in single organs only. (**E**) A few examples of previously unreported rare signatures. SBS signatures can be viewed at https://signal.mutationalsignatures.com/explore/study/6?mutationType=1.

substitution (TBS) signature. Exploring all triple-base possibilities, we obtain an 864-channel profile that systematically reports an excess of TTT>AAA followed by TTT>GAA, TTT>CAA, and minor contributions of TTG>AAA and TTC>AAA. We propose that this is called TBS1 (Fig. 3E and table S27). However, the numbers of mutations and implicated samples are too low to perform a formal mutational signature analysis.

Our curation steps uncovered several DBS signatures, including previously reported ones, that comprise adjacent substitutions that are not in cis and are simply the mathematical outcome of an associated SBS hypermutator (fig. S8). For example, DBS3 and DBS10 were similar and correlated with polymerase ε (*POLE*)–attributed SBS10a (fig. S8, A to C). In silico analysis showed that a DBS pattern that recapitulates DBS3 and DBS10 could be reproduced from hypermutated samples of SBS10a (fig. S8D). The alleged double substitutions were not, in fact, in cis. Similarly, DBS12 (associated with SBS105), DBS14 (associated with

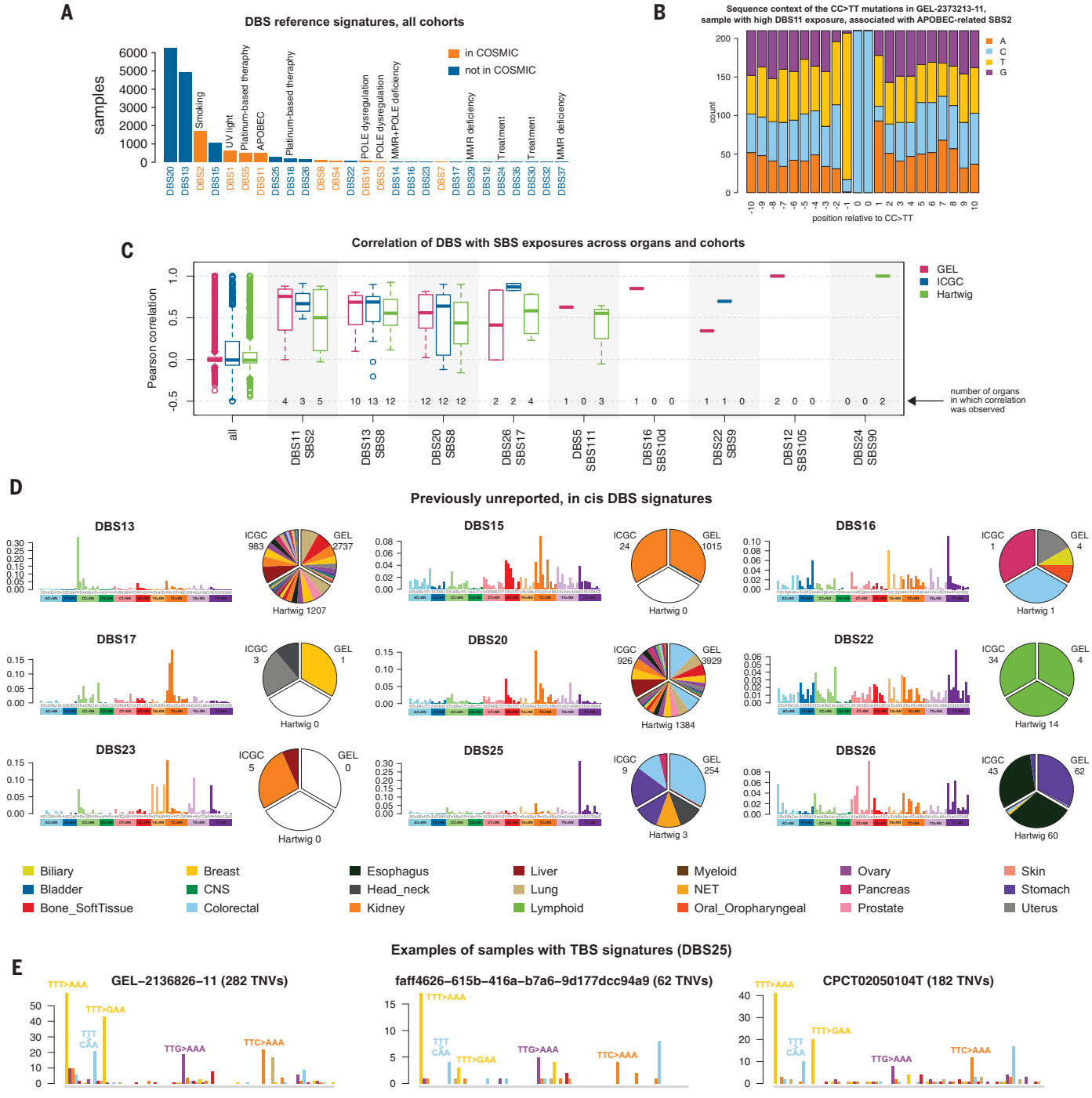


Fig. 3. DBS signatures across the cohort. (**A**) Frequency of DBS signatures in the present study. (**B**) Flanking sequence context surrounding mutated dinucleotides of DBS11, which is correlated with APOBEC-related SBS2, to demonstrate a preference for the TpCCpN context, similar to the TpCpN sequence predilection of APOBECs. (**C**) Correlation of DBSs with SBS exposures across cohorts. Numbers in each column report the number of organs implicated in the correlative analyses. A correlation is computed independently for each organ. Correlations are displayed as a box plots, which denote median (horizontal line) and 25th to 75th percentiles (boxes). The lower and upper whiskers span 1.5 times the interquartile range (1.5 × IQR). (**D**) Examples of previously unreported DBS signatures. (**E**) Samples with TBS1. The total numbers of samples and triple-nucleotide variants are too low to perform a formal mutational signature analysis. All DBS signatures identified in the present study can be viewed at https://signal.mutationalsignatures.com/explore/study/6?mutationType=2.

SBS14), DBS29 (associated with SBS20), and DBS37 (associated with SBS26) could all be generated mathematically from their associated SBS signatures (fig. S8, E to H), indicating that these were not true dinucleotides but simply single-nucleotide variants occurring next to each other by chance. One exception, DBS24—associated with SBS90, attributed to duocarmycin exposure—has a pattern that can be mostly recapitulated by simulation of SBS90, apart from the CT>AA component (fig. S8I). Three signatures (DBS23, DBS32, and DBS35) were not

in the GEL cohort and could not be curated, owing to a lack of access to sequencing data.

Contrasting previously unreported signatures with previously reported endogenous processes

Deamination and amplified deamination

Pervasive patterns of deamination are widely observed in malignant and nonmalignant tissues. SBS1, characterized by C>T mutations at CpG, is due to deamination of methylcytosine, whereas SBS2 and SBS13 are due to APOBEC-related deamination. Both are likely physiological: SBS1 occurs by natural hydrolytic processes, whereas SBS2 and SBS13 arise through transient single-stranded DNA availability (*20*).

Two rare signatures also characterized by C>T transitions at CpG are SBS96 and SBS95, differing by their ability to demonstrate marked hypermutator phenotypes and relative C>T peak heights (Fig. 4, A and B). SBS96, present in 18 of 12,222 GEL samples [0.15%, table S23; reported as due to inherited and/or acquired mutations in *MBD4* (*21*)], has C>T at ACG as its tallest peak. We identified germline-truncating *MBD4* mutations with loss of heterozygosity (LOH) of the alternative allele to explain 12 of 18 samples (6 of 10 patients) with SBS96 (table S28 and Fig. 4C). *MBD4* germline variants were also seen in 35 other GEL patients, yet SBS96 was not observed in their tumors because the wild-type parental allele was intact in all assessable cases. Notably, SBS96 was observed in extremely rare cancers such as myxofibrosarcomas and uveal melanoma. SBS95 is distinguishable from SBS96 by having its tallest peak at CCG and by exhibiting transcriptional strand bias (TSB). SBS95 occurred in a lymphoid and a stomach cancer in the GEL cohort and one head and neck cancer in the ICGC cohort (table S23). None had *MBD4* mutations. The cause for SBS95 remains unclear.

Two signatures were characterized by C>N at CpG (Fig. 4A). SBS87 (*22*), with its tallest peak at CCG, was observed in one breast cancer. A related signature with C>N at all CpGs, SBS105, was reported in one breast and one bladder cancer in the GEL dataset. Although we have not found a cause for SBS105, it is associated with DBS12, a mathematical outcome of a high rate of SBS105 (fig. S8E), and does not exhibit TSB. Mechanistically, SBS105 would require deamination at CpGs followed by generic misincorporation during DNA replication and/or repair, not limited to the A rule (*23*), to generate this pattern.

Even though all of these signatures occur at CpGs, they have distinguishing characteristics. Discriminating *MBD4*-related SBS96 is particularly important given reports that such tumors have sensitivities to checkpoint therapies (*24*).

DNA repair deficiency phenomena

A multitude of DNA repair genes and proteins serve as guardians of the genome (*25*). If compromised, they can result in mutational patterns in human cells.

Compromised components of base excision repair

SBS18 was previously described in neuroblastomas and adrenocortical cancers (*5*, *26*). Subsequently, a hypermutated version of a signature similar to SBS18 was described in tumors from patients with biallelic mutations in *MUTYH*, a gene that encodes a base excision repair (BER) protein (MUTYH glycosylase) that corrects oxidative damage (*27*). Recently, it was demonstrated that OGG1 (8-oxo-guanine glycosylase) loss produces a phenocopy of SBS18 and that signatures defined by tall peaks at C>A at GCA, ACA, GCT, and TCT are due to an excess of 8-oxo-dG (*25*). Signature SBS108 resembles SBS18 and could be due to 8-oxo-dG (*25*), although it has differences such as inclusion of the tallest C>A peak at GCA instead of TCT (Fig. 4D). Notably, three GEL patients who have tumors with SBS108 carried a germline polymorphism in *OGG1* (rs113561019 p.G308E) that has been reported as a risk allele in microsatellite-stable hereditary nonpolyposis colorectal cancer (*28*). We assessed the background frequency of this allele and found it present in 98 individuals (0.85%, table S28). Fifteen patients had tumors estimated as homozygous for the rs113561019 allele, including the 3 with SBS108 and 12 additional samples. It is possible that the presence of other strong signatures encumbered the detection of SBS108 in these cases.

Seven samples from six patients carried SBS30 associated with variants in *NTHL1*, another BER glycosylase (Fig. 4E and tables S29 and S30). Two cases had germline nonsense *NTHL1* mutations with associated loss of the wild-type parental allele. Three cases had somatic rearrangements that resulted in deletion of large sections of the gene. One of the three, the ovarian cancer GEL-2126555-11, had a mixed phenotype of SBS30 and features of BRCA2 loss and carried a germline *BRCA2* frameshift mutation that creates deletion signatures. This case also had two somatic deletions that affect *NTHL1*.

Mismatch repair and polymerase abnormalities

Replication of the nuclear genome occurs with high fidelity because of postreplicative mismatch repair (MMR) activity and proofreading capacity of DNA polymerases—particularly *POLE* and *POLD*. Unsurprisingly, MMR pathway defects and selected mutations in polymerases cause high rates of mutagenesis.

We confirm four previously reported MMR deficiency (MMRd) signatures: SBS6, SBS15, SBS26, and SBS44 (Fig. 4, F and G). As noted previously (*5*, *9*, *14*), we find a particular enrichment of mutations in MMR genes (*MLH1*, *MSH2*, and *MSH6*) in SBS6, SBS15, and SBS44, many of which exhibit loss of the alternative parental allele as well (Fig. 4H and tables S29 and S30). In SBS26, previously shown to be identical to signatures of human knockouts of *PMS2* (*25*), we indeed identified 14 *PMS2*-inactivating mutations (10 germline and 4 somatic, 7 of 14 biallelic) in 23 samples from 22 patients (Fig. 4H and tables S29 and S30). Some caution should be exercised in interpreting somatic mutations in cancers with high burdens of substitutions or indels, as these could be chance events. Regardless, it is noteworthy that a genetic driver cannot be identified for approximately one of every two cancers with MMRd signatures. Methylation data are not available for assessment.

In addition, we confirm that SBS10a is associated with POLE dysregulation. All 65 GEL samples with SBS10a had *POLE* mutations consistent with proofreading dysfunction (Fig. 4H and tables S29 and S30). We also confirm that two of five GEL samples with SBS10d carried the previously reported *POLD1* exonuclease domain mutation p.(Asp316Asn) (*29*). Here we report an identical p.(Asp903Tyr) mutation in DNA polymerase domain B in the remaining three samples.

Two signatures—SBS14 (MMRd and POLE dysfunction) and SBS20 (MMRd and POLD dysfunction)—were previously attributed to a mixed phenotype of MMRd and polymerase mutants (*29*). Of 14 samples with SBS14, 13 had potential *POLE* drivers (four established and nine putative; tables S29 and S30). Of the same 14 samples, 11 also had truncating mutations in MMR genes (*MSH6*, *MSH2*, *MLH1*, or *PMS2*; 3 germline and 15 somatic mutations), but only 6 appeared to be inactivated on both parental alleles. Similarly, of eight samples with SBS20, four had missense drivers in *POLD1* (one germline and four somatic). Seven of the eight also had inactivating mutations in *MSH6* or *MSH2* [germline (n = 4) and/or somatic (n = 7)], six of which showed biallelic inactivation. Again, all of these tumors had high mutation burdens; thus, some mutations could be chance events due to high MMRd mutation rates. Moreover, elevated mutation rates of MMRd signatures cause a high likelihood of substitutions occurring adjacent to each other, falsely creating DBS patterns DBS14, DBS29, and DBS37 (fig. S8, F to H).

Lastly, we identify a signature with a defined C>T peak at GCG, SBS97, most closely resembling SBS15; however, it can be distinguished from SBS15 by strong T>C at GTC and T>G at GTT trinucleotides (Fig. 4F). Observed in just three colorectal cancers in the GEL dataset and five in the HMF dataset, SBS97 is rare and has a strong hypermutator phenotype

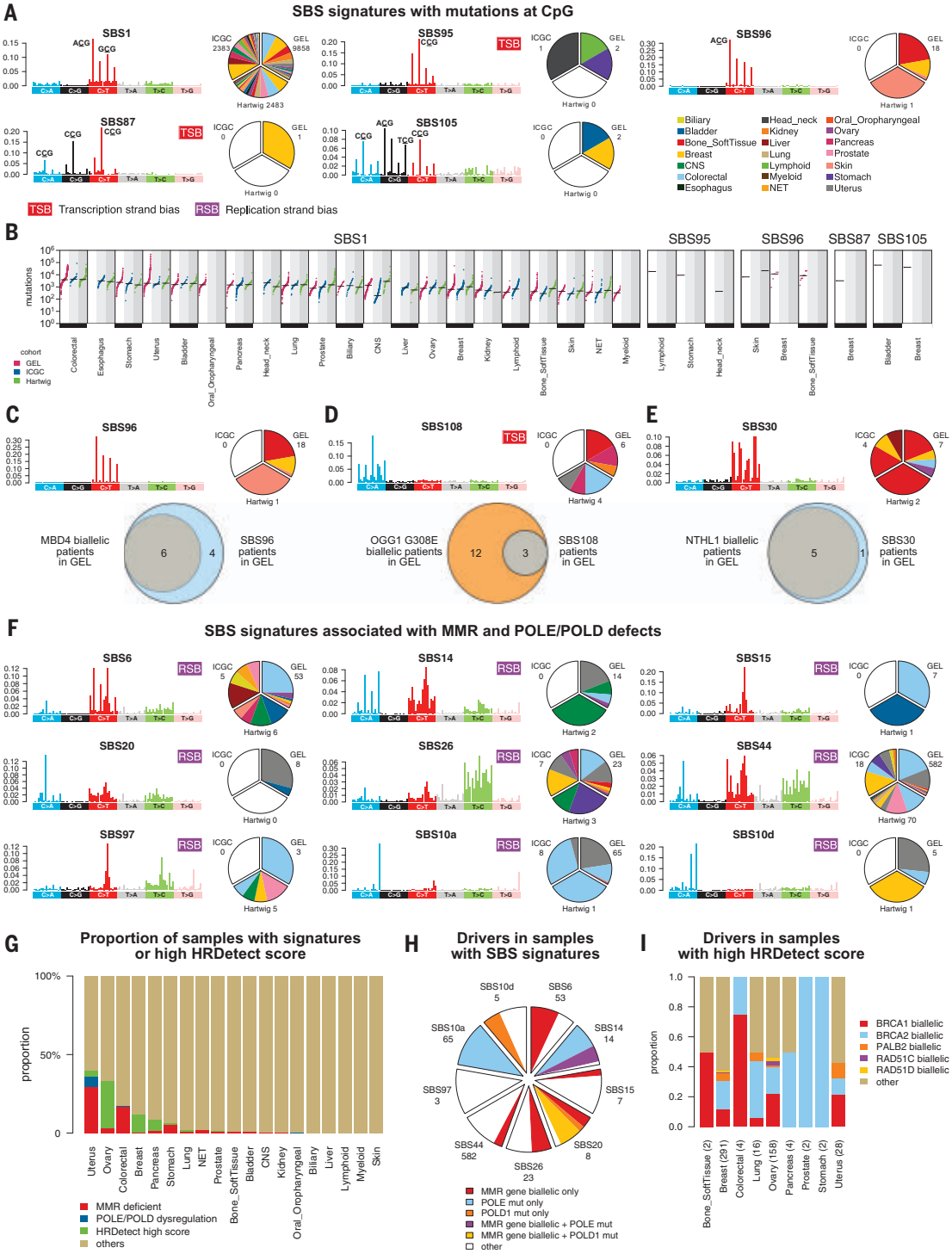


Fig. 4. Signatures associated with endogenous mutational processes. (**A**) Five signatures characterized by substitutions at NCG nucleotides are contrasted with each other. SBS profiles are shown, along with prevalence of signatures across three cohorts (GEL, ICGC, and HMF), TSB, and RSB. (**B**) Mutation burdens associated with each signature in all tumor types. SBS1 is common and seen in all cohorts in many tumor types. SBS95, SBS96, SBS87, and SBS105 are rare and associated with a higher mutation burden than that of SBS1. The *y* axis shows the mutation count on a log scale. Summaries comparing signatures in etiological groupings can be found at https://signal.mutationalsignatures.com/explore/study/6. (**C**) SBS96 signature and prevalence among samples in the three cohorts. The Venn diagram illustrates the number of patients in the GEL cohort with SBS96 and biallelic loss of *MBD4*. (**D**) SBS108 signature and prevalence among samples in the three cohorts. The Venn diagram illustrates the number of patients in the GEL cohort with SBS108 and a biallelic *OGG1* Gly^{308}-Glu mutation. (**E**) SBS30 signature and prevalence among samples in the three cohorts. The Venn diagram illustrates the number of patients in the GEL cohort with SBS30 and biallelic loss of *NTHL1*. Only samples with SBS30 were inspected for *NTHL1* mutations. (**F**) SBSs associated with MMR and *POLE* and *POLD* gene defects. (**G**) Proportion of samples across GEL organs with signatures related to MMR or *POLE* and *POLD* defects or with high HRDetect scores. (**H**) Proportion of samples with MMR biallelic loss and/or POLE and POLD dysregulation in each group of samples presenting an SBS mutational signature in (F). (**I**) Proportion of HRDetect-high samples with biallelic loss in genes linked to homologous recombination repair. Numbers in parentheses indicate numbers of samples.

[29 to 65 substitutions per megabase (subs/Mb)] and an even higher indel rate (67 to 99 indels/Mb). All three GEL cases also have considerable structural variation (0.02 to 0.05 variations/Mb), revealing that chromosomal instability and microsatellite instability are not mutually exclusive in colorectal cancer. No causative drivers have been confirmed to date.

In all, MMRd and polymerase-dysregulated signatures are prominent in colorectal (413 of 2348, ~18%) and uterine (258 of 713, 36%) cancers in the GEL cohort (Fig. 4D). Sporadic incidences of MMRd occurred in stomach (11), prostate (3), pancreas (1), ovary (18), NETs (2), lung (8), kidney (9), oropharyngeal (1), CNS (3), breast (14), sarcoma (16), and bladder (3) cancers (total 89 of 9161, <1% total), with clinical implications.

Compromised components of double-strand break repair

SBS3 was previously shown to distinguish *BRCA1*- and *BRCA2*-null cancers from sporadic breast cancers (*6*). SBS8 is increased in *BRCA1*- and *BRCA2*-null cancers (*9*). We applied a previously developed algorithm, HRDetect (*17*, *30*), designed to detect tumors with BRCA1- and BRCA2-compromised double-strand break repair, to the GEL cohort (Fig. 4G, fig. S9, and table S31). The prevalence of HRDetect high scores (5th- to 95th-percentile confidence interval above 0.5) was variable within each tumor type. More than 30% of all ovarian cancers had high HRDetect scores; ~11% of breast cancers (predominantly estrogen receptor–positive cancers), ~7% of pancreatic cancers, ~4% of uterine cancers, 1.6% of lung cancers, ~1% of stomach cancer, and <1% of prostate, bone, and colorectal cancers also had high scores. The causes of high HRDetect scores were identified in 231 of 493 individuals (47%, biallelic loss confirmed in 40%; Fig. 4I and tables S29 and S30) and included germline and somatic mutations in *BRCA1*, *BRCA2*, *PALB2*, *RAD51C*, and *RAD51D*, as described previously (*6*, *9*, *31*, *32*). Promoter hypermethylation data were not available.

Environmental sources of mutational signatures

UV-like C>N signatures at CCN and TCN

We reinforce SBS7a (defined by C>T at CCN and TCN) in skin tumors with associated DBS1 characterized by CC>TT dinucleotides (*33*) (Fig. 5, A and B). However, we highlight three signatures that occurred at similar trinucleotides (CCN and TCN) and that could be misinterpreted as related to UV light exposure but may be due to alternative etiologies.

SBS129, observed once in a nodular malignant melanoma (GEL-2501934-11) and once in a leiomyosarcoma (GEL-2300438-11), is characterized by C>T transitions at CCN trinucleotides (particularly CCA and CCT) but not TCN trinucleotides (Fig. 5A). It is distinguishable from SBS7a by its rarity and lack of CC>TT dinucleotides. However, like SBS7a, SBS129 presents transcriptional strand asymmetry with excess C>T mutations on the nontranscribed strand. Apart from somatic *TP53* mutations, no other potential genetic associations have been identified.

SBS38 is identical in its trinucleotide preponderance to SBS129, except it results in C>A transversions instead (Fig. 5A). Although this signature has been reported before (*14*), it is rare, and its etiology is unknown. Here, we identify it in 30 cancers (29 skin, 1 lung; table S23) in the GEL cohort and note that it can either be a dominating phenotype or occur in combination with SBS7a, SBS17, and SBS18. Notably, among the samples affected by SBS38, we found all three anorectal mucosal cancers—which are aggressive, unusual mucosal melanocytic cancers—in the GEL cohort. The occurrence of the uncommon signature SBS38 in a very rare tumor type hints at a germline genetic predisposition, yet we have not been able to identify a causative gene. Minor TSB is noted with more mutations on the transcribed strand for C>A mutations.

Lastly, SBS137 was observed twice in GEL brain cancers (table S23) and would superficially seem highly similar to UV-associated signatures (Fig. 5A). Critically, affected tumors do not have a CC>TT DBS signature (Fig. 5B) and demonstrate TSB in the opposite direction of that associated with exposure to UV light (table S32), with an excess of C>T mutations on the transcribed strand (likely representing an excess of G>A on the nontranscribed strand). The tallest peak of SBS137 is at CCC, unlike the most prominent SBS7a peak at TCC. By contrast, the classic appearance of SBS7a and DBS1 is observed in a metastatic CNS lesion derived from a cutaneous primary cancer (GEL-2906789-11), which suggests that SBS137 is a distinct signature of currently uncertain cause.

Aristolochic acid exposure and similar patterns

SBS22 is due to exposure to aristolochic acid (AAI) (*33*) (Fig. 5C). In the GEL dataset, all three renal cancers with SBS22 were from patients who reported ethnic-minority ancestry. None reported past exposure to AAI.

We noted that SBS113 is similar to SBS22; has tall peaks in T>A with additional contributions from T>C at GTN; and is seen in one CNS (GEL-2585923-11), one colorectal (GEL-2282347-11), and one lung (GEL-2158956-11) cancer. There is no history of exposure to AAI in these patients, although all three patients had complex therapeutic histories, including extensive exposure to psychotropic drugs and antiepileptics.

In previous work, alternative compounds from unrelated chemical families [specifically dibenzo[a,l]pyrene (DBP) and its diol epoxide (DBPDE) from the polycyclic aromatic hydrocarbons family in tobacco smoke] that caused bulky adducts on adenines similar to AAI were capable of generating signatures nearly identical to those associated with AAI (*33*). Thus, given its similarities to SBS22, SBS113 may represent mutational processes with alternative etiologies that also cause adducted adenines.

Platinum exposure

SBS31 is associated with prior platinum exposure (*34*) (Fig. 5D). This signature—characterized by C>T peaks at CCC and CCT; C>A peaks at ACC, CCT, and GCC; and a modest T>A peak at CTN—has previously been demonstrated experimentally in a human cell line model (*33*).

SBS35 is similar to SBS31, though it has smaller contributions at all trinucleotides (*14*). SBS104 may be related to SBS31, as it shows C>A peaks at CCC and CCT and was found in two HMF metastatic samples that had exposure to platinum. Two additional signatures, SBS111 and SBS112, have the components seen in SBS31, albeit with additional features, particularly in C>A and noisier C>T components. Clinical histories of the patients carrying these signatures reveal that all had past diagnoses of primary malignant neoplasms of the ovary, stomach, esophagus, or breast, or of non-Hodgkin's lymphoma, and presented with secondary or new primary malignancies. All patients had complex chemotherapy that involved exposure to platinum. Perhaps these signatures are complex outcomes of multiple treatments and immune modulation on the genome of the tumor samples isolated for sequencing. Two DBS platinum signatures (DBS5 and DBS18) are also associated with these SBS signatures (Fig. 5E).

Tobacco-related signatures and others with similar C>A components

SBS4, associated with tobacco smoke exposure (*33*) (Fig. 5F), is seen mainly in lung cancers (at high levels of ~90 subs/Mb). SBS4 is noted very rarely in other tumor types (table S23): It was observed in just one breast cancer (GEL-2791664-11), one colorectal lesion noted to be metastatic (GEL-2842602-11), one diffuse astrocytoma (GEL-2645293-11), and two CNS lesions of unknown primary origin (GEL-2860373-11 and GEL-2500813-11). The presence of SBS4 is supported by DBS2 (Fig. 5G) and TSB in all of these cases and probably indicates metastatic lesions of primary lung cancer in these instances.

Two signatures that have similarities to SBS4 are SBS94 and SBS109 (Fig. 5F). SBS94 is characterized by C>A mutations with the tallest peak at CCC followed by CCA. In colon (nine cases) and breast (one case) cancers, this signature does not have a hypermutator phenotype nor an associated DBS, but TSB and RSB are noted for C>A variants (table S19).

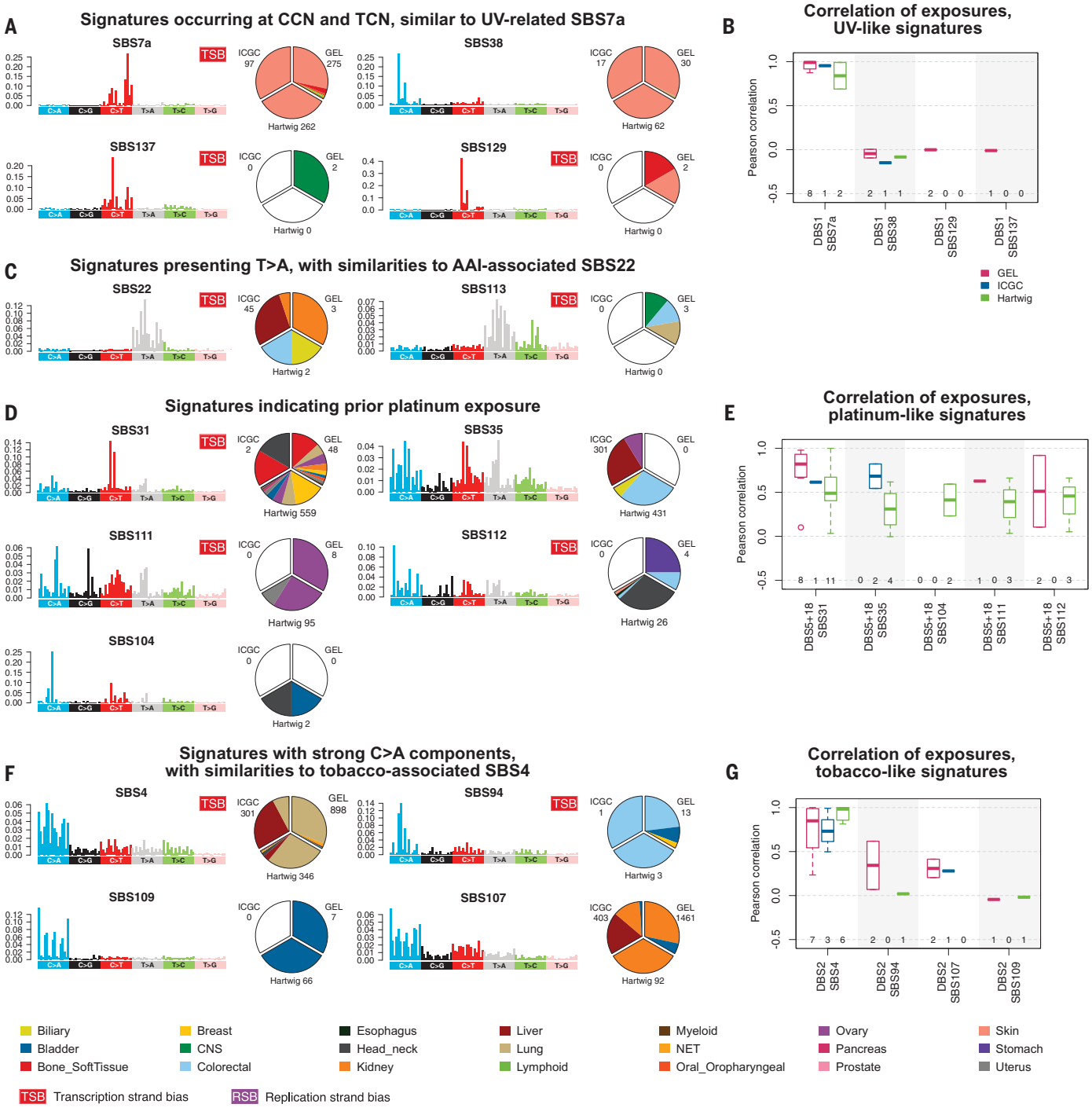


Fig. 5. Signatures associated with environmental mutational processes. (**A**) SBS signatures occurring at CCN and TCN, with similarities to UV-related SBS7a. (**B**) Correlations of the signatures in (A) with DBS signatures. (**C**) SBS signatures presenting T>A, with similarities to AAI-associated SBS22. (**D**) SBS signatures indicating prior platinum exposure. (**E**) Correlations of the signatures in (D) with DBS signatures. (**F**) SBS signatures presenting C>A with similarities to tobacco-associated SBS4. (**G**) Correlation of the signatures in (F) with DBS signatures. The number in each column indicates the number of organs implicated in the correlative analyses. A correlation is computed independently for each organ, and correlations are displayed as a box plot. Box plots denote median (horizontal line) and 25th to 75th percentiles (boxes). The lower and upper whiskers extend 1.5 × IQR.

In bladder cancers (three cases), there is a marked DBS pattern despite low mutational burden (0.15 to 8 subs/Mb). The cause for this curious difference in tissue behavior is unclear.

SBS109 is a C>A pattern with tall peaks at NCA and NCT, although the tallest peaks are primarily located at ACA and TCT. Only seven bladder cancers demonstrate this phenotype, and it does not have any associated DBS or TSB. The mutation burden is also low, at only 0.3 to 3 subs/Mb. SBS107 is seen at low levels in bladder and kidney cancers (0.04 to

6 subs/Mb) across many samples of these tumor types. It is a common signature in kidney and bladder cancers (1461/1704) and is akin to SBS109 but with additional contributions at NCC.

Multiple signatures have been attributed to environmental exposures, but we will not examine all of them. Signature not discussed here include SBS11 (associated with alkylation on a mismatch repair–deficient background), SBS90 (associated with duocarmycin), and SBS88 (reported as due to colibactin produced by pks^{+}*E. coli* infection) (*35*, *36*).

Future use of mutational signatures

The ever-increasing number of mutational signatures presents the challenge of using mutational signature analysis in practice, whether in a new study of aggregated samples or for individual patients. To address this obstacle, we acknowledge that most nonexpert users will aim to understand which mutational signatures are present in a new set of patient samples that are often tissue specific. This "signature-fitting" process will require users to ask which predefined signatures of a circumscribed set are present in their samples. To explore how to better perform this fitting process, we first consider mutational signatures per tumor type, using CNS tumors from the GEL cohort (Fig. 6) as an example. Additional per tumor signature summaries can be found in figs. S10 to S51 and at https://signal.mutationalsignatures.com/explore/study/6.

Per–tumor-type summaries

A total of 809 whole-genome–sequenced CNS tumors have been evaluated. In the GEL dataset, 6% of CNS tumors have rare signatures (Fig. 6, A and B). Previously reported common signatures in the GEL CNS cohort include age-associated SBS1 and SBS5, HR-deficiency–related SBS3 and SBS8. A previously unreported common signature, SBS120, is present in many CNS tumors at a low to moderate mutation rate (Fig. 6C). Common CNS signatures exhibit clear and reproducible tissue specificity (fig. S52). Previously reported rare signatures observed in the GEL CNS cohort include APOBEC signatures SBS2 and SBS13, SBS17 of unknown etiology, SBS11 due to temozolomide on an MMR-deficient genetic background, and MMRd signatures (SBS14) (Fig. 6D). We noted rare occurrences of tobacco-related SBS4 and UV-induced SBS7a in metastatic lesions.

We also identified several previously unreported rare signatures in CNS tumors (Fig. 6E). These include the aforementioned signature SBS113, which has similarities to AAI-related SBS22. SBS121, defined by C>G at ACT and TCT, is common in colorectal and stomach cancers but seen in only three CNS tumors, and its etiology is unknown. SBS119 is present in a single CNS tumor as a hypermutator phenotype (28 subs/Mb) in the GEL cohort and in two CNS tumors in the HMF cohort. Lastly, SBS137 is distinct from signatures associated with UV exposure, has no DBS despite a high mutational burden, and is CNS specific and rare.

DBS1 and DBS2 are associated with exposure to UV light and tobacco smoke, respectively, and are seen in the samples with SBS7a and SBS4. Three previously unreported DBS signatures are observed (Fig. 6F): DBS13 and DBS20 are relatively common, whereas DBS14 is rare and is due to the high mutational burden of MMRd-associated signature SBS14 (fig. S8F).

Reassuringly, common signatures are prevalent in all three cohorts (GEL, ICGC, and HMF) (Fig. 6, G and H), whereas the presence of rare signatures is a function of the size of the examined cohort. In all, this example highlights the landscape of common and rare signatures in this tumor type (Fig. 6G) and provides pointers regarding how to pragmatically use mutational signatures for signature fitting of new samples.

Fitting signatures: FitMS

Cancer samples have a median of five common signatures; when rare signatures are present, there is usually only one instance per sample (fig. S53, A and B). Learning from these results, we developed a signature-fitting algorithm, Signature Fit Multi-Step (FitMS) (fig. S53C), which first estimates the presence of tissue-relevant common signatures and then attempts to identify additional rare signatures in a second step, assuming that only one or two rare signatures may be present.

To evaluate the performance of FitMS, we performed a simulation study in which each simulated sample comprised five organ-specific common signatures and some samples carried one rare signature (materials and methods). We contrasted three strategies: (i) fitting all common and rare signatures together in a single step (fit all); (ii) a two-step method fitting common signatures using a constraint of positive residuals that are matched to rare signatures in the second step (constrainedFit); and (iii) a two-step method fitting common signatures, followed by the addition of rare signatures one at a time to achieve a reduction in the residual between true and modeled catalogs (errorReduction). The two-step errorReduction FitMS strategy demonstrated superior performance (fig. S53, D to F), improving the fit of common and rare signatures better than the constrainedFit or fit all approaches. Moreover, using organ-specific common signatures rather than corresponding reference signatures improved the accuracy of signature assignment (fig. S53, G to I).

Therefore, for practical purposes, to assess which signatures are present in any new sample or set of samples, we recommend this two-step process (Fig. 7): first fitting common organ-specific signatures followed by a search for rare signatures, which can be achieved using FitMS.

Discussion

We report a comprehensive SBS and DBS signatures analysis of a large cohort of 18,640 whole-genome–sequenced tumors. Notably, most of these samples were from patients recruited via the UK NHS (12,222) from across England, and the availability of open-access WGS cancer data from ICGC and HMF were crucial for validation of findings. In all, 40 SBS and 18 DBS signatures that had not been previously reported were revealed as a result of the increase in WGS cohort size. We were also able to confirm 42 previously reported SBS signatures and 9 previous DBS signatures. We introduce the notion of common and rare signatures for each tumor type and observe that although the cohort of whole-genome–sequenced cancers has increased substantially, most of the common signatures have been identified, and many of the previously unreported signatures are low-frequency, rare processes. The landscape of signatures is thus likely to be saturating.

The power to accurately discern mutational signatures with a pure WGS dataset is orders of magnitude greater than that obtained by means of other sequencing strategies. The genomic footprint for whole-exome sequencing (WES) is 100-fold lower and is 2000- to 4000-fold lower in targeted sequencing (TS) experiments. Analyzing only whole-genome–sequenced cancers rather than pooling data from diverse sequencing strategies also avoids issues related to differing AT or GC representation in WES or TS data, which influence signature extractions.

Methodologically, several points are noteworthy. First, grouping samples by organs and focusing on common mutational profiles has produced signatures that are highly reproducible across cohorts. Removing atypical samples in the first extraction step is especially important for large cohorts, in which very rare signatures may be present and could interfere with the accurate identification of common signatures. Second, the use of three large independent cohorts is crucial for validation of signatures found in single organs, such as SBS120, and that could otherwise be mistaken for other signatures or considered artifactual. Third, although some signatures may have 96-element SBS profiles that are very similar to those of other well-known signatures, additional information, such as co-occurrence with DBS signatures or TSB and/or RSB, can suggest a different etiology and help validate them as distinct signatures. Thus, deeper investigation can often show distinctions that indicate diverse etiologies, a caveat that must be considered when using mutational signatures in future analyses.

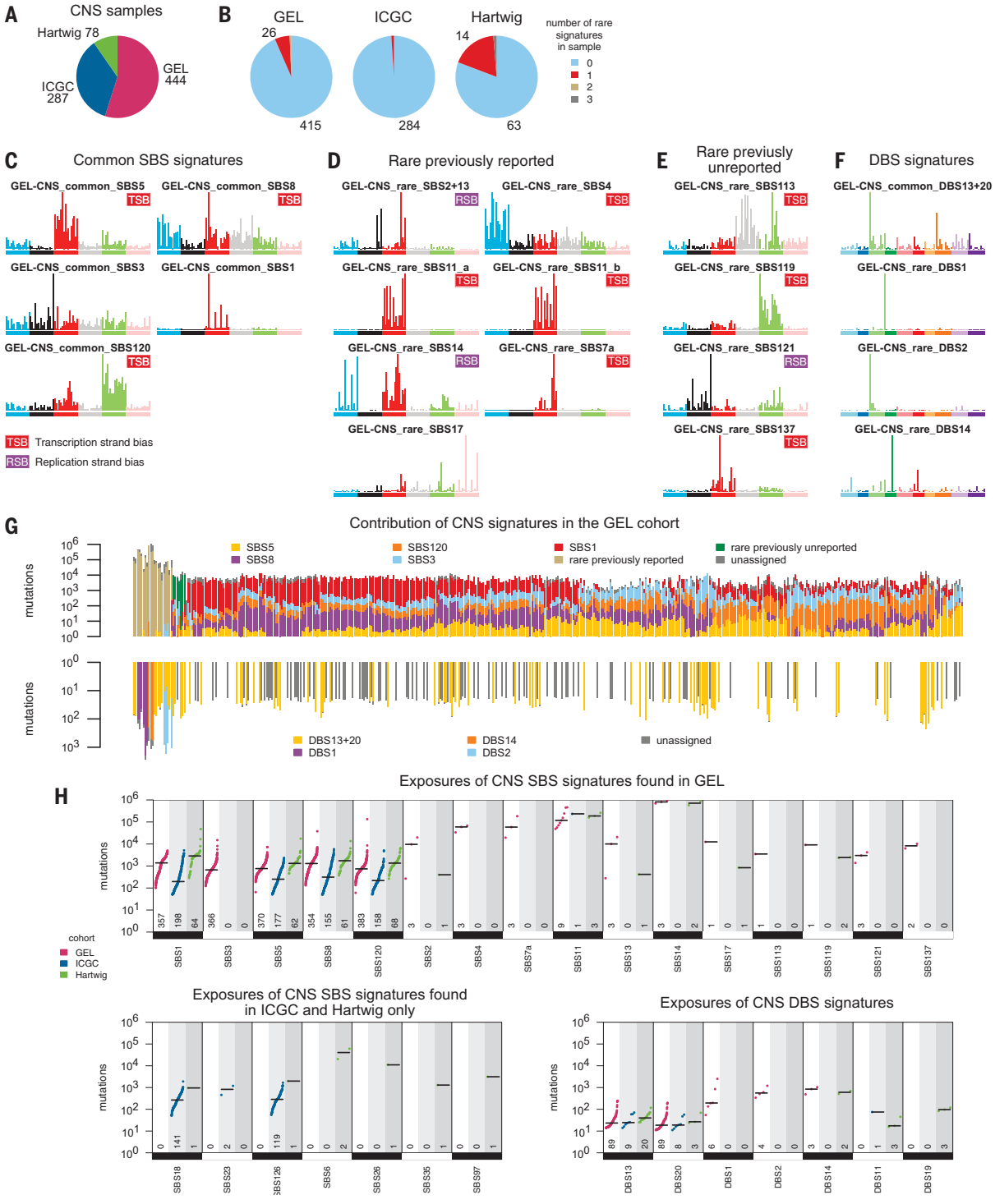


Fig. 6. Summary of SBS signatures and DBS signatures in CNS tumors from the GEL dataset. (**A**) CNS tumors from GEL, ICGC and HMF cohorts. (**B**) Most CNS tumors have common signatures only (light blue), and a fraction have one rare signature (maroon). In the pie charts, numbers <5% are not shown. (**C**) Common SBS signatures in CNS GEL tumors. (**D**) Previously reported rare SBS signatures in CNS GEL tumors. (**E**) Previously unreported rare SBS signatures in CNS GEL tumors. (**F**) DBS signatures in CNS GEL tumors. (**G**) Distribution of mutational signatures in all CNS GEL tumors. For each sample, the total number of mutations is shown in a log scale, whereas signature exposure proportions are colored linearly. Samples are clustered according to the exposure proportions using hierarchical clustering with average linkage. (**H**) Mutational frequencies of common and rare signatures of CNS GEL cancers. Numbers at the bottom indicate the numbers of samples with each signature.

From a biological perspective, it is essential to distinguish signatures that provide diagnostic insights or that are therapeutically informative from other signatures, particularly when they share feature similarities. Notable examples deliberated earlier include distinguishing *MBD4*-compromised SBS96 from other signatures with CpG propensity or correctly differentiating signatures that occur predominantly at CCN and TCN from UV-related SBS7a.

Additionally, we highlight endogenous signatures indicative of pathway defects that are

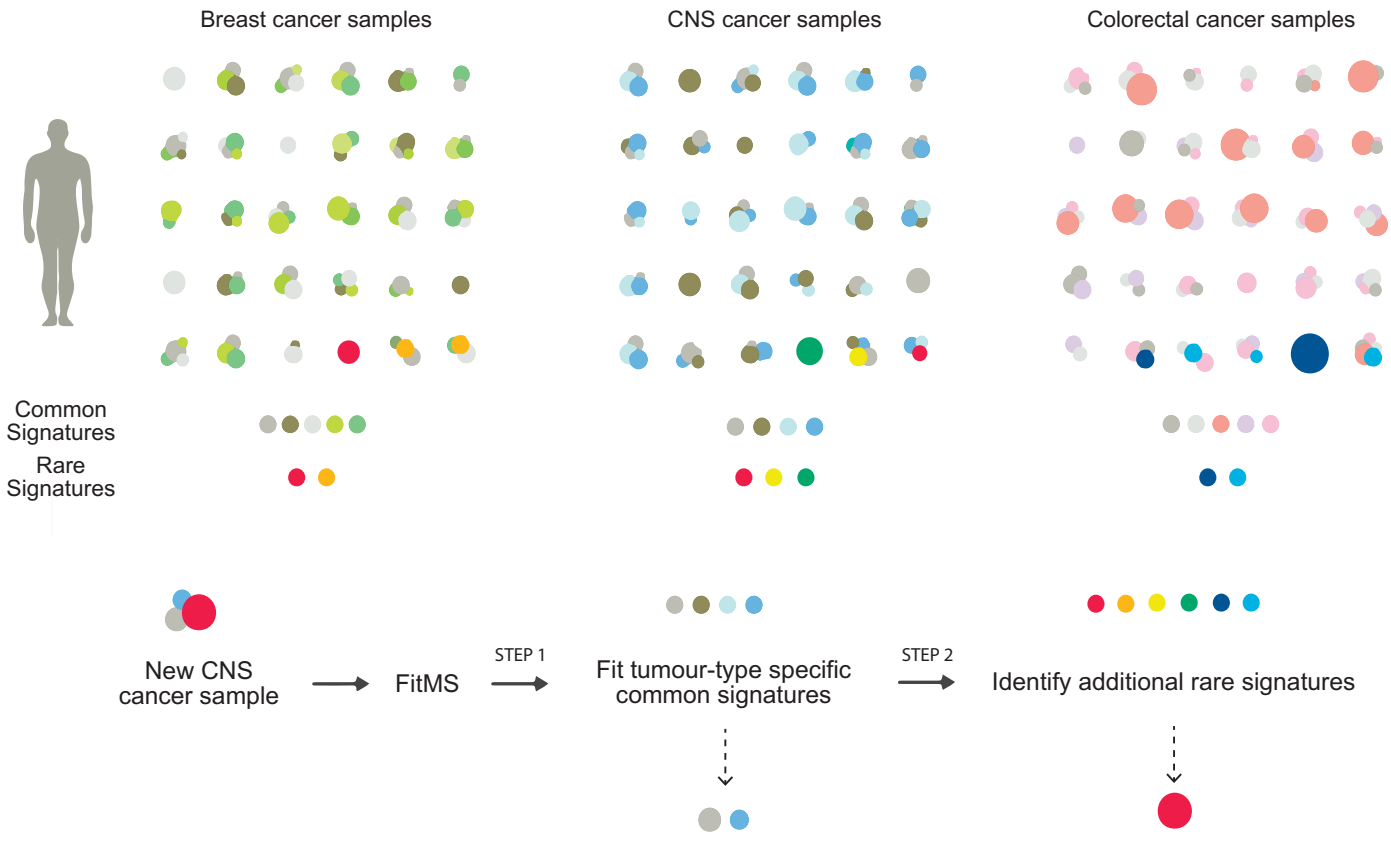


Fig. 7. Illustration of common and rare mutational signatures in cancer samples and the workflow of FitMS. Schematic depiction of common (gray and lighter colors) versus rare signatures in three example tumor types (breast, CNS, and colorectal cancers). Each patient could have different amounts of some (or all) of the common signatures. Occasionally, a patient may carry a rare signature as well (bright colors). Some common signatures are ubiquitous and present in nearly all tumor types, whereas other common signatures are restricted to certain tumor types. Rare signatures may be unique (yellow dot) or may also occur in other tumor types (red dots). We propose a practical package, FitMS, that makes use of the insights obtained in this study. Given a new sample—for example, a new brain cancer WGS mutation catalog—FitMS will fit common CNS signatures before attempting to discover additional rare signatures seen in CNS and other tumors.

detectable using WGS signatures but for which a genetic driver cannot be identified. It is noteworthy that a causal genetic event could not be detected for one in two cases with MMRd and one in two cases with HR deficiency, indicating that signature analysis is more sensitive for identifying these defects than examination of mutations in selected genes. Furthermore, an agnostic WGS approach to tumor characterization will help reveal abnormalities that we currently neither seek nor detect using customary diagnostic pathways. For example, we found MMRd-associated signatures with a frequency of <1% in many tumor types, including stomach, prostate, pancreas, ovary, NET, lung, kidney, oropharyngeal, CNS, breast, sarcoma, and bladder cancers. Given the reported therapeutic relationships between MMRd phenotypes and immune checkpoint inhibitors (*37*–*39*), from a personalized pan-cancer therapeutics perspective, many of these patients may be eligible for treatment options that would otherwise not be available to them.

At present, many of the previously unreported signatures have no known etiology. This is not surprising because of the complexity of drawing causal relationships, particularly for endogenous signatures, which can be the outcome of multiple co-occurring events. For example, a gene defect in *MBD4* could convert the ubiquitous C>T at CpG into a hypermutator phenotype (SBS96), or a pathophysiological state such as replication stress could amplify APOBEC-related SBS13. Some endogenous signatures may only manifest as part of an adaptive response to stressful stimuli. For example, SBS17, defined by T>G and T>C mutations, was reported in mouse cells that have been through immortalization, in normal human cells treated with 5-FU (*40*), and in a wide variety of cancers. Thus, many of the signatures of unknown etiology could be due to not just a single-gene defect but multigene or complex pathway abnormalities (*41*) and/or may become overt after an adaptive response to cellular stress. Further work will be required to fully comprehend the causes of many cancer mutational signatures.

As our knowledge base increases, the complexity of assigning genetic causality to signatures is evident in examples such as the *OGG1* polymorphic risk allele, for which some patients exhibit SBS108 clearly but others do not. Looking forward, alternative strategies may be needed to detect the contribution of moderate- and lower-penetrance germline risk alleles to somatic signatures in large cohorts.

Notably, the present analysis introduces the concept of common versus rare signatures within each tumor type. It highlights how an increased number of samples may help discern common signatures that occur at low levels for specific tumor types. Greater sample numbers may also help unveil signatures that occur at a low frequency in the population. Crucially, the availability of independent, open-access datasets such as those from the ICGC and HMF has been instrumental in corroborating these common and rare signatures identified within the GEL dataset. Although it is far simpler to discuss signatures as unifying reference patterns across all organs, it is important to note that these are mathematical reference patterns, an average of many extractions, and not necessarily an accurate biological

representation of the process in any given tissue. For users seeking to learn what signatures may be present in a new set of samples, it may be more advisable to use organ-specific signatures rather than mathematically averaged signatures to perform an analysis.

Thus, we suggest a strategy of using mutational signatures, which accounts for the biological insights and complexities described in this work. FitMS invites the user to use common organ-specific signatures in the first instance, followed by a hunt for rare signatures (Fig. 7).

Indeed, as many national cancer genomic endeavors get underway worldwide over the next decade, we look forward to applying WGS data maximally to advance individualized cancer care.

Materials and methods summary

Datasets

We considered three large pan-cancer whole-genome cohorts: the GEL version 8 cohort of the 100,000 Genomes Project (*7*), comprising 15,838 whole-genome–sequenced paired sample; the ICGC cohort (*9*, *11*), comprising 3001 whole-genome–sequenced paired samples; and the HMF cohort (*12*), comprising 3417 whole-genome–sequenced tumor samples. After considering comparability of tumor types across cohorts and QC of GEL data, we focused our analysis on 12,222 high-quality WGS GEL cases (tables S1 to S6).

Mutational signature extraction

For each tumor sample, we counted the number of somatic mutations and constructed SBS (96-channel) and DBS (78-channel) mutational catalogs (tables S7 and S8). Mutational signatures were analyzed independently for each tumor type in each of the three cohorts (Fig. 1C). First, we clustered mutational catalogs and excluded samples with unusual profiles (hierarchical clustering using 1 – cosine similarity as distance) (fig. S1, A to C), aiming at reducing the number of rare, complicating signatures and obtaining fewer, more accurate signatures. Second, we used non-negative matrix factorization with Kullback-Leibler divergence (KLD) optimization, repeated bootstrapping (at least 300 bootstraps), and removed poor local minima (*17*). We identified a set of common mutational signatures that were organ and cohort specific. Third, we fitted the common signatures into all samples of a given cohort and tissue type and identified samples with high reconstruction error to identify unexplained processes or rare mutational signatures (details in supplementary materials; fig. S1, D to H, and tables S9 to S12).

To define signature exposures in each sample, we used a signature fit procedure. Briefly, the number of mutations attributed to each signature in each sample was estimated from organ-specific signatures detected in their originating cohort using KLD optimization (non-negative linear models R package) and bootstrapping (200 bootstraps) (*17*). Point estimates of exposures were the median of the exposures obtained from bootstrapping. Exposures below 5% of the total SBS burden or below 25% of DBS burden per sample were set to zero because of the risk of overfitting.

Reference signatures

To permit comparability across cohorts and organs, we defined "reference signatures" (Fig. 1G). In brief, we clustered all common and rare mutational signatures (757 SBS or 301 DBS signatures) (tables S13 and S14) and obtained clusters of highly similar signatures (187 SBS and 60 DBS clusters). Cluster averages were termed "distinct patterns" (tables S15 and S16). We assigned each distinct pattern to one of three groups: (i) a reliably recurrent distinct pattern observed in multiple independent extractions, (ii) a mix of two or more distinct patterns, (iii) a singleton pattern found in one organ in one cohort (tables S17 and S18). Recurrent distinct patterns were additionally clustered to remove patterns that may simply be a variant of another pattern. Mixed distinct patterns that could be estimated as a combination of two distinct patterns using non-negative least squares were dismissed. Singleton distinct patterns were also curated and dismissed if they could simply be variants of other reference signatures. If they had been reported in other studies, they were retained as reference signatures. A total of 120 SBS and 39 DBS reference signatures were identified.

A QC status—green, amber, or red—was assigned to each of the reference signatures. QC green signatures were those extracted independently multiple times and/or reported in orthogonal studies. QC amber status was given to signatures with limited supporting evidence, such as signatures identified in only one extraction and not reported previously. QC red status was assigned to signatures that were mathematical or alignment artifacts. After QC, 82 of 120 SBS and 27 of 39 DBS reference signatures remained QC green [tables S19 and S20, SBS/DBS final reference signatures (tables S21 and S22), exposures (tables S23 and S24)]. Conversion matrices that map reference signatures to organ-specific signatures of each cohort are in tables S25 and S26.

Additional analytics related to correlations with germline and somatic driver events can be found in the supplementary materials and tables S27 to S30.

RSB and TSB were calculated as in previous work (*42*). Briefly, we counted classes of single-nucleotide variants (C>A, C>G, C>T, T>A, T>C, and T>G) taking into account whether they appeared on the lagging or leading strand (according to MCF-7 reference Repli-Seq data) or on the transcribed or nontranscribed strand (according to gene orientation) (*42*). A paired two-tailed Student's t test was used to determine the significant deviation from the natural bias given by the region's base content. The log2 ratio was used to determine the size of the asymmetry between the two strands (table S32).

HRDetect scores were computed as previously described (*17*, *30*). HRDetect input features are exposures of SBS3 and SBS8, proportions of short deletions at microhomology, HRD-LOH index, and exposures of rearrangement signatures 3 and 5. Rearrangement signature exposures were estimated by means of KLD optimization, bootstrapping, and previously published rearrangement signatures (*17*). HRDetect scores were computed both as point estimates and also as a distribution obtained from 1000 bootstrapped scores, as previously described (*17*) (table S31).

FitMS and simulation study

Signature Fit Multi-Step (FitMS) is an algorithm designed to estimate signature exposures, taking advantage of the concept of common and rare signatures. FitMS has two steps. In the first step, only common signature exposures are estimated. In the second step, the presence of potential rare signatures is estimated, achievable through two possible strategies: constrainedFit or errorReduction. The constrainedFit strategy uses constrained nonnegative least squares (limSolve R package) to estimate the residual between the observed and reconstructed catalogs, using only common signatures. If this residual resembled a rare signature (cosine similarity of at least 0.8) then we assumed that a rare signature was present in the sample. In the errorReduction strategy, the error (KLD) between the original catalog and the fit obtained using only common signatures was compared with the error obtained using one additional rare signature, for all rare signatures considered. A rare signature is considered present if the reduction in error is at least 15%. Regardless of strategy, we recomputed sample exposures using both common signatures and any additional rare signatures.

To evaluate the performance of FitMS (fig. S53), we simulated 100 genomes, each containing five common signatures chosen randomly from the nine common SBS breast cancer signatures in the GEL dataset. In addition, one rare signature was added to 25 of 100 samples, and each rare signature was chosen randomly from 54 possible rare, curated SBS reference signatures observed in at least two independent extractions. We compared the two FitMS strategies against a "fit-all" strategy in which the aforementioned 9 common breast cancer signatures and 54 rare, curated

signatures were used in one single signature-fitting process. Each signature-fit strategy produced a first estimate of the exposures, which tended to overfit signatures into samples, resulting in false-positive assignments of signatures to samples with very few associated mutations. To remove false positives, we removed signature exposures that represented a very small proportion of mutations, testing thresholds from 0 to 10% of total sample mutations (fig. S53, D to I).

For users of FitMS, the set of common and rare signatures that could be fitted into any sample is thus organ dependent. Lists of signatures per organ can be found in table S33.

Full materials and methods are available in the supplementary materials (*43*).

REFERENCES AND NOTES

1. H. Sung *et al*., Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **71**, 209–249 (2021). doi: 10.3322/caac.21660; pmid: 33538338
2. M. R. Stratton, P. J. Campbell, P. A. Futreal, The cancer genome. *Nature* **458**, 719–724 (2009). doi: 10.1038/nature07943; pmid: 19360079
3. D. R. Bentley *et al*., Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008). doi: 10.1038/nature07517; pmid: 18987734
4. T. Helleday, S. Eshtad, S. Nik-Zainal, Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.* **15**, 585–598 (2014). doi: 10.1038/nrg3729; pmid: 24981601
5. L. B. Alexandrov *et al*., Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013). doi: 10.1038/nature12477; pmid: 23945592
6. S. Nik-Zainal *et al*., Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012). doi: 10.1016/j.cell.2012.04.024; pmid: 22608084
7. C. Turnbull, Introducing whole-genome sequencing into routine cancer care: The Genomics England 100 000 Genomes Project. *Ann. Oncol.* **29**, 784–787 (2018). doi: 10.1093/annonc/mdy054; pmid: 29462260
8. J. Ma, J. Setton, N. Y. Lee, N. Riaz, S. N. Powell, The therapeutic significance of mutational signatures from DNA repair deficiency in cancer. *Nat. Commun.* **9**, 3292 (2018). doi: 10.1038/s41467-018-05228-y; pmid: 30120226
9. S. Nik-Zainal *et al*., Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016). doi: 10.1038/nature17676; pmid: 27135926
10. C. Ganini *et al*., Global mapping of cancers: The Cancer Genome Atlas and beyond. *Mol. Oncol.* **15**, 2823–2840 (2021). doi: 10.1002/1878-0261.13056; pmid: 34245122
11. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020). doi: 10.1038/s41586-020-1969-6; pmid: 32025007
12. P. Priestley *et al*., Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019). doi: 10.1038/s41586-019-1689-y; pmid: 31645765
13. E. Turro *et al*., Whole-genome sequencing of patients with rare diseases in a national health system. *Nature* **583**, 96–102 (2020). doi: 10.1038/s41586-020-2434-2; pmid: 32581362
14. L. B. Alexandrov *et al*., The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020). doi: 10.1038/s41586-020-1943-3; pmid: 32025018
15. A. R. J. Lawson *et al*., Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science* **370**, 75–82 (2020). doi: 10.1126/science.aba8347; pmid: 33004514
16. S. M. A. Islam *et al*., Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *bioRxiv*2020.2012.2013.422570 [Preprint] (2021); doi: 10.1101/2020.12.13.422570
17. A. Degasperi *et al*., A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies. *Nat. Cancer* **1**, 249–263 (2020). doi: 10.1038/s43018-020-0027-5; pmid: 32118208
18. O. Pich *et al*., The mutational footprints of cancer therapies. *Nat. Genet.* **51**, 1732–1740 (2019). doi: 10.1038/s41588-019-0525-5; pmid: 31740835
19. P. S. Robinson *et al*., Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. *Nat. Genet.* **53**, 1434–1442 (2021). doi: 10.1038/s41588-021-00930-y; pmid: 34594041
20. C. Swanton, N. McGranahan, G. J. Starrett, R. S. Harris, APOBEC Enzymes: Mutagenic Fuel for Cancer Evolution and Heterogeneity. *Cancer Discov.* **5**, 704–712 (2015). doi: 10.1158/2159-8290.CD-15-0344; pmid: 26091828
21. M. A. Sanders *et al*., MBD4 guards against methylation damage and germ line deficiency predisposes to clonal hematopoiesis and early-onset AML. *Blood* **132**, 1526–1534 (2018). doi: 10.1182/blood-2018-05-852566; pmid: 30049810
22. B. Li *et al*., Therapy-induced mutations drive the genomic landscape of relapsed acute lymphoblastic leukemia. *Blood* **135**, 41–55 (2020). doi: 10.1182/blood.2019002220; pmid: 31697823
23. B. S. Strauss, The 'A rule' of mutagen specificity: A consequence of DNA polymerase bypass of non-instructional lesions? *BioEssays* **13**, 79–84 (1991). doi: 10.1002/bies.950130206; pmid: 2029269
24. M. Rodrigues *et al*., Outlier response to anti-PD1 in uveal melanoma reveals germline MBD4 mutations in hypermutated tumors. *Nat. Commun.* **9**, 1866 (2018). doi: 10.1038/s41467-018-04322-5; pmid: 29760383
25. X. Zou *et al*., A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. *Nat. Cancer* **2**, 643–657 (2021). doi: 10.1038/s43018-021-00200-0; pmid: 34164627
26. C. Pilati *et al*., Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas. *J. Pathol.* **242**, 10–15 (2017). doi: 10.1002/path.4880; pmid: 28127763
27. A. Viel *et al*., A Specific Mutational Signature Associated with DNA 8-Oxoguanine Persistence in MUTYH-defective Colorectal Cancer. *EBioMedicine* **20**, 39–49 (2017). doi: 10.1016/j.ebiom.2017.04.022; pmid: 28551381
28. P. Garre *et al*., Analysis of the oxidative damage repair genes NUDT1, OGG1, and MUTYH in patients from mismatch repair proficient HNPCC families (MSS-HNPCC). *Clin. Cancer Res.* **17**, 1701–1712 (2011). doi: 10.1158/1078-0432.CCR-10-2491; pmid: 21355073
29. N. J. Haradhvala *et al*., Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nat. Commun.* **9**, 1746 (2018). doi: 10.1038/s41467-018-04002-4; pmid: 29717118
30. H. Davies *et al*., HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* **23**, 517–525 (2017). doi: 10.1038/nm.4292; pmid: 28288110
31. P. Polak *et al*., A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat. Genet.* **49**, 1476–1486 (2017). doi: 10.1038/ng.3934; pmid: 28825726
32. J. Staaf *et al*., Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. *Nat. Med.* **25**, 1526–1533 (2019). doi: 10.1038/s41591-019-0582-4; pmid: 31570822
33. J. E. Kucab *et al*., A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**, 821–836.e16 (2019). doi: 10.1016/j.cell.2019.03.001; pmid: 30982602
34. E. Pleasance *et al*., Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes. *Nat. Cancer* **1**, 452–468 (2020). doi: 10.1038/s43018-020-0050-6; pmid: 35121966
35. C. Pleguezuelos-Manzano *et al*., Mutational signature in colorectal cancer caused by genotoxic pks$^+$ E. coli. *Nature* **580**, 269–273 (2020). doi: 10.1038/s41586-020-2080-8; pmid: 32106218
36. P. J. Dziubańska-Kusibab *et al*., Colibactin DNA-damage signature indicates mutational impact in colorectal cancer. *Nat. Med.* **26**, 1063–1069 (2020). doi: 10.1038/s41591-020-0908-2; pmid: 32483361
37. D. T. Le *et al*.., Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* **357**, 409–413 (2017). doi: 10.1126/science.aan6733; pmid: 28596308
38. A. Marabelle *et al*., Efficacy of Pembrolizumab in Patients With Noncolorectal High Microsatellite Instability/Mismatch Repair-Deficient Cancer: Results From the Phase II KEYNOTE-158 Study. *J. Clin. Oncol.* **38**, 1–10 (2020). doi: 10.1200/JCO.19.02105; pmid: 31682550
39. H. Veeraraghavan *et al*., Machine learning-based prediction of microsatellite instability and high tumor mutation burden from contrast-enhanced computed tomography in endometrial cancers. *Sci. Rep.* **10**, 17769 (2020). doi: 10.1038/s41598-020-72475-9; pmid: 33082371
40. S. Christensen *et al*., 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. *Nat. Commun.* **10**, 4571 (2019). doi: 10.1038/s41467-019-12594-8; pmid: 31594944
41. G. Rospo *et al*., Evolving neoantigen profiles in colorectal cancers with DNA repair defects. *Genome Med.* **11**, 42 (2019). doi: 10.1186/s13073-019-0654-6; pmid: 31253177
42. S. Morganella *et al*., The topography of mutational processes in breast cancer genomes. *Nat. Commun.* **7**, 11383 (2016). doi: 10.1038/ncomms11383; pmid: 27136393
43. See supplementary materials.
44. A. Degasperi *et al*., Mutational signatures in whole-genome-sequenced cancers in the UK population, Mutational Signatures Data. v1, Zenodo (2022), https://doi.org/10.5281/zenodo.5571551.
45. A. Degasperi *et al*., Mutational signatures in whole-genome-sequenced cancers in the UK population, Supplementary Code S1 and S2. v1, Zenodo (2022), https://doi.org/10.5281/zenodo.5570307.

ACKNOWLEDGMENTS

This work was enabled by access to data and findings generated by the 100,000 Genomes Project, under the auspices of the Pan-Cancer GeCIP (project RR239). The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care) funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK, and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support. This publication and the underlying research are facilitated by data that were generated by the Hartwig Medical Foundation (HMF) and the Center for Personalized Cancer Treatment (CPCT) in the Netherlands, as well as the International Cancer Genome Consortium. **Funding:** This work was supported by Cancer Research UK (CRUK) Advanced Clinician Scientist Award grant C60100/A23916, Dr. Josef Steiner Cancer Research Award 2019, Medical Research Council (MRC) Grant-in-Aid to the MRC Cancer Unit, CRUK Pioneer Award C60100/A23433, CRUK Early Detection Project Award C60100/A27815, CRUK Grand Challenge Award grant C60100/A25274, and NIHR Research Professorship NIHR301627. This work was also supported by National Institute of Health Research (NIHR) Cambridge Biomedical Research Centre grant BRC-125-20014. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. **Author contributions:** Conceptualization: S.N.-Z. and A.D. Methodology: A.D., T.D.A., H.R.D., and A.M.-M. Resources, new genomics, and clinical data: M.A.B. and Genomics England Research Consortium. Software: A.D., X.Z., T.D.A., A.M.-M., J.M.L.D., S.S., J.C., and D.P.-G. Data curation: Genomics England Research Consortium, A.D., S.N.-Z., H.R.D., A.M.-M., Y.M., and T.D.A. Investigation: S.N.-Z., A.D., X.Z., T.D.A., H.R.D., A.M.-M., G.C.C.K., J.M.L.D., L.H., L.C., G.R., V.Y.W.W., A.S.N., A.B., S.E.M., J.Y., D.P.-G., Y.M., and C.B. Visualization: A.D. Funding acquisition: S.N.Z. Project administration: S.N.-Z. Supervision: S.N.-Z. and H.R.D. Writing – original draft: S.N.-Z. and A.D. Writing – review and editing: S.N.-Z., A.D., G.C.C.K., H.R.D., X.Z., and S.S. **Competing interests:** A.D., X.Z., H.R.D., and S.N.-Z. hold patents or have submitted applications on clinical algorithms of mutational signatures [MMRDetect (PCT/EP2022/057387), HRDetect (PCT/EP2017/060294), clinical use of signatures (PCT/EP2017/060289), rearrangement signature methods (PCT/EP2017/060279), clinical predictor (PCT/EP2017/060298), and hotspots for chromosomal rearrangements (PCT/EP2017/060298] and, during this project, have served in advisory roles for AstraZeneca, Artios Pharma, and the Scottish Genomes Project. **Data and materials availability:** Primary data from the 100,000 Genomes Project, which are held in a secure research environment, are available to registered users. See https://www.genomicsengland.co.uk/research/academic for further information or contact M.A.B., Chief Scientific Officer at Genomics England (matt.brown@genomicsengland.co.uk). The ICGC cohort contains 2471 cancer whole genomes from PCAWG (EGAS00001001692) and 530 additional breast cancer

genomes (450 from EGAS00001001178 and 80 from EGAD00001002740). The HMF cohort can be accessed via at www.hartwigmedicalfoundation.nl/en. Data access requests and institutional agreements are required for all cohorts. The results of the analysis can be browsed at https://signal.mutationalsignatures.com/explore/study/6 or downloaded as a compressed archive from Zenodo (*44*). The code used for this analysis is available as Code S1 (R scripts) and Code S2 (new version of R package *signature.tools.lib* that includes FitMS) on Zenodo (*45*).

Genomics England Research Consortium
J. C. Ambrose[1], P. Arumugam[1], R. Bevers[1], M. Bleda[1], F. Boardman-Pretty[1,2], C. R. Boustred[1], H. Brittain[1], M. J. Caulfield[1,2], G. C. Chan[1], T. Fowler[1], A. Giess[1], A. Hamblin[1], S. Henderson[1,2], T. J. P. Hubbard[1], R. Jackson[1], L. J. Jones[1,2], D. Kasperaviciute[1,2], M. Kayikci[1], A. Kousathanas[1], L. Lahnstein[1], S. E. A. Leigh[1], I. U. S. Leong[1], F. J. Lopez[1], F. Maleady-Crowe[1], M. McEntagart[1], F. Minneci[1], L. Moutsianas[1,2], M. Mueller[1,2], N. Murugaesu[1], A. C. Need[1,2], P. O'Donovan[1], C. A. Odhams[1], C. Patch[1,2], D. Perez-Gil[1], M. B. Pereira[1], J. Pullinger[1], T. Rahim[1], A. Rendon[1], T. Rogers[1], K. Savage[1], K. Sawant[1], R. H. Scott[1], A. Siddiq[1], A. Sieghart[1], S. C. Smith[1], A. Sosinsky[1,2], A. Stuckey[1], M. Tanguy[1], A. L. Taylor Tavares[1], E. R. A. Thomas[1,2], S. R. Thompson[1], A. Tucci[1,2], M. J. Welland[1], E. Williams[1], K. Witkowska[1,2], S. M. Wood[1,2]

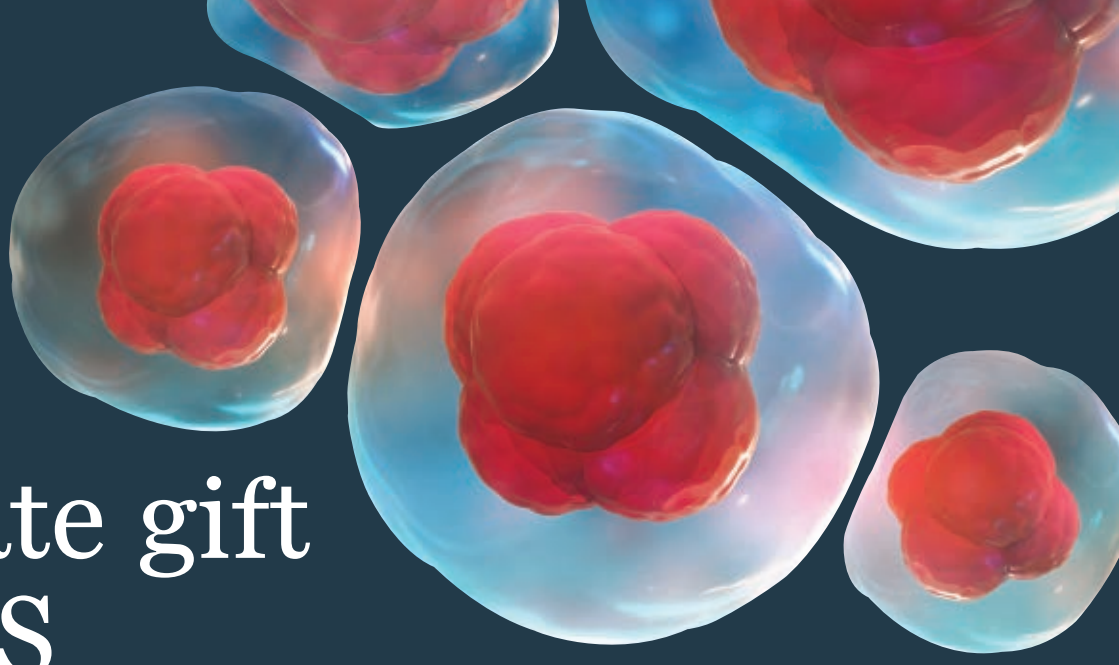
[1]Genomics England, Queen Mary University of London, Dawson Hall, Charterhouse Square, London EC1M 6BQ, UK. [2]William Harvey Research Institute, Queen Mary University of London, London EC1M 6BQ, UK.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abl9283
Materials and Methods
Figs. S1 to S53
Tables S1 to S33
References (*46*, *47*)

17 August 2021; accepted 4 March 2022
10.1126/science.abl9283

1848 SOCIETY
AAAS



An estate gift to AAAS

Going all the way back to 1848, our founding year, the American Association for the Advancement of Science (AAAS) has been deeply committed to advancing science, engineering and innovation around the world for the benefit of all people.

Today, we are dedicated to advocating for science and scientific evidence to be fully and positively integrated into public policy and for the community to speak with one voice to advance science and engineering in the United States and around the world.

By making AAAS a beneficiary of your will, trust, retirement plan or life insurance policy, you will become a member of our 1848 Society and will help fuel our work on behalf of science and society – including publishing the world's most promising, innovative research in the *Science* family of journals and engaging in the issues that matter locally, nationally and around the world.

"As a teacher and instructor, I bear responsibility for the younger generations. If you have extra resources, concentrate them on organizations, like AAAS, that are doing work for all."

—Prof. Elisabeth Ervin-Blankenheim, 1848 Society member

If you intend to include AAAS in your estate plans, provide this information to your lawyer or financial adviser:

Legal Name: American Association for the Advancement of Science
Federal Tax ID Number: 53-0196568
Address: 1200 New York Avenue, NW, Washington, DC 20005

If you would like more information on making an estate gift to AAAS, cut out and return the form below or send an email to philanthropy@aaas.org. Additional details are also available online at **www.aaas.org/1848Society**.

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

cut here

Yes, I would like more information about joining the AAAS 1848 Society.

1848 SOCIETY
AAAS

PLEASE CONTACT ME AT:

Name: _____

Address: _____

City: _____ State: _____ Zip code: _____ Country: _____

Email: _____ Phone: _____

RETURN THIS FORM TO:
AAAS Office of Philanthropy and Strategic Partnerships • 1200 New York Avenue, NW • Washington, DC 20005 USA

RESEARCH ARTICLE SUMMARY

CORONAVIRUS

Structural basis for potent antibody neutralization of SARS-CoV-2 variants including B.1.1.529

Tongqing Zhou*†, Lingshu Wang†, John Misasi*†, Amarendra Pegu, Yi Zhang, Darcy R. Harris, Adam S. Olia, Chloe Adrienna Talana, Eun Sung Yang, Man Chen, Misook Choe, Wei Shi, I-Ting Teng, Adrian Creanga, Claudia Jenkins, Kwanyee Leung, Tracy Liu, Erik-Stephane D. Stancofski, Tyler Stephens, Baoshan Zhang, Yaroslav Tsybovsky, Barney S. Graham, John R. Mascola‡, Nancy J. Sullivan‡, Peter D. Kwong*‡

INTRODUCTION: The emergence and rapid spread of the B.1.1.529 (Omicron) variant of concern (VOC) with 37 mutations in the spike protein has raised alarm. Especially troublesome are the 15–amino acid substitutions in the receptor binding domain (RBD) because RBD-directed antibodies have been the only antibodies found to retain sufficient potency against other variants. To identify antibodies that effectively neutralize B.1.1.529, we evaluated RBD-directed antibodies for their ability to bind and neutralize B.1.1.529 and determined their modes of recognition using functional assays and cryo–electron microscopy (cryo-EM) structures.

RATIONALE: The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) B.1.1.529 VOC is substantially resistant to neutralization by most monoclonal antibodies and by vaccinee and convalescent sera. Identifying monoclonal antibodies that retain neutralization potency against this variant and understanding their structural mechanism of recognition should inform the development of vaccines and antibody therapeutics that maintain effectiveness.

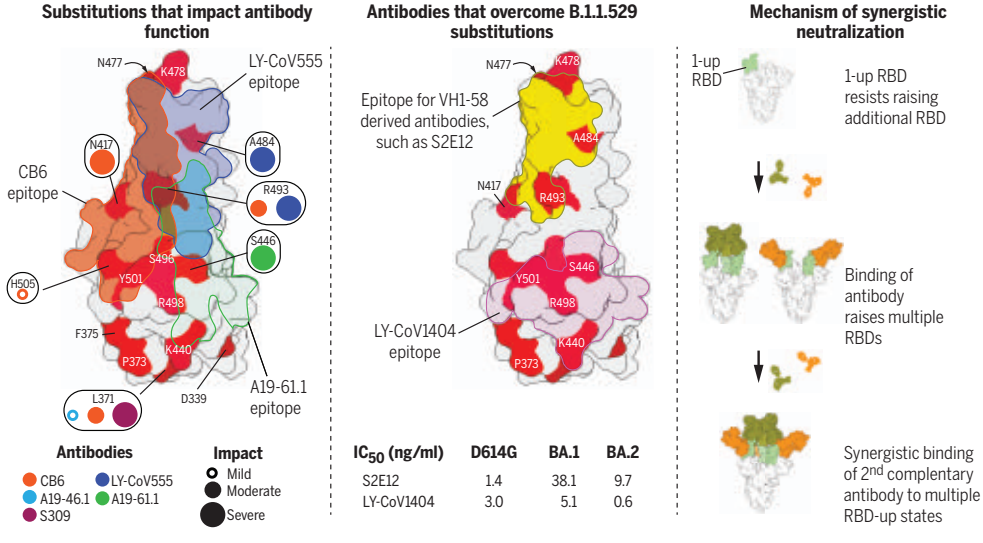
RESULTS: The cryo-EM structure of the B.1.1.529 spike in its prefusion conformation revealed a single RBD-up conformation, with RBD substitutions localized to the outer surface of the spike. Despite this localization, RBD substitutions directly contacted or bordered epitopes of all previously identified RBD-directed neutralizing antibodies. Our studies revealed antibodies A23-58.1, B1-182.1, COV2-2196, S2E12, A19-46.1, S309, and LY-CoV1404 to nevertheless maintain substantial neutralization against this emerging variant. To provide structural and functional explanations, we determined cryo-EM structures of antibody-spike complexes and used virus particles representing each of the 15 single-amino-acid substitutions in RBD to delineate their functional impact. For class I and II antibodies that compete with angiotensin-converting enzyme 2 (ACE2) for binding to spike, such as VH1-58–derived antibodies B1-182.1 and S2E12, these analyses revealed potent neutralization to require smaller antibody side chains that accommodate the S477N mutation. For others, such as LY-CoV555 and A19-46.1, the epitopes of these antibodies bordered multiple RBD substitutions. Both E484A or Q493R greatly reduced binding for LY-CoV555, whereas for A19-46.1, these substitutions were generally tolerated, with the cryo-EM structure of A19-46.1 and spike revealing a two-RBD-up conformation and A19-46.1 binding only to up RBDs. For class III and IV antibodies that bind outside of the ACE2-binding surface—such as A19-61.1, COV2-2130, S309, and LY-CoV1404—individual B.1.1.529 substitutions were generally tolerated. However, A19-61.1 neutralization was eliminated by G446S; COV2-2130 showed substantially lower neutralization of B.1.1.529, but no single mutation exhibited a substantial impact; S309 retained potency against B.1.1.529, although not against the BA.2 sublineage variant (half-maximal inhibitory concentration reduced to 1374 ng/ml); and LY-CoV1404 retained potent neutralization (5.1 and 0.6 ng/ml against BA.1 and BA.2 sublineages, respectively). Last, we assessed combinations of monoclonal antibodies and found several, including the combination of B1-182.1 and A19-46.1, to show neutralization synergy. The structure of the ternary complex of B.1.1.529 spike with these two antibodies suggested the induction of the preferred up-RBD binding conformation by B1-182.1 to facilitate cooperative binding by A19-46.1 as the basis for their synergy.

CONCLUSION: Although Omicron mutations cluster, they nonetheless affect virtually all known RBD-directed neutralizing antibodies. Our study reveals the structural basis by which select RBD-directed antibodies such as S2E12 and LY-CoV1404 retain potent neutralization of B.1.1.529. We further identified antibody combinations that can be used for treatment and have demonstrated how these combinations overcome extensive spike mutations. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: tzhou@nih.gov (T.Z.); john.misasi@nih.gov (J.M.); pdkwong@nih.gov (P.D.K.)
†These authors contributed equally to this work.
‡These authors contributed equally to this work.
Cite this article as T. Zhou *et al.*, *Science* **376**, eabn8897 (2022). DOI: 10.1126/science.abn8897

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abn8897

Structural basis for potent neutralization of B.1.1.529 by monoclonal antibodies and their combination. The RBD molecular surface is shown, colored by B.1.1.529 substitutions in red. (Left) Example epitopes are highlighted, along with the impact of each amino acid substitution. (Middle) Epitopes are shown for antibodies S2E12 and LY-CoV1404, which retain potent neutralization of B.1.1.529. (Right) Antibody combinations with complementary recognition modes can bind cooperatively to mediate synergistic neutralization.



RESEARCH ARTICLE

CORONAVIRUS

Structural basis for potent antibody neutralization of SARS-CoV-2 variants including B.1.1.529

Tongqing Zhou[1]*†, Lingshu Wang[1]†, John Misasi[1]*†, Amarendra Pegu[1], Yi Zhang[1], Darcy R. Harris[1], Adam S. Olia[1], Chloe Adrienna Talana[1], Eun Sung Yang[1], Man Chen[1], Misook Choe[1], Wei Shi[1], I-Ting Teng[1], Adrian Creanga[1], Claudia Jenkins[2], Kwanyee Leung[1], Tracy Liu[1], Erik-Stephane D. Stancofski[1], Tyler Stephens[2], Baoshan Zhang[1], Yaroslav Tsybovsky[2], Barney S. Graham[1], John R. Mascola[1]‡, Nancy J. Sullivan[1]‡, Peter D. Kwong[1]*‡

The rapid spread of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) B.1.1.529 (Omicron) variant and its resistance to neutralization by vaccinee and convalescent sera are driving a search for monoclonal antibodies with potent neutralization. To provide insight into effective neutralization, we determined cryo–electron microscopy structures and evaluated receptor binding domain (RBD) antibodies for their ability to bind and neutralize B.1.1.529. Mutations altered 16% of the B.1.1.529 RBD surface, clustered on an RBD ridge overlapping the angiotensin-converting enzyme 2 (ACE2)–binding surface and reduced binding of most antibodies. Substantial inhibitory activity was retained by select monoclonal antibodies—including A23-58.1, B1-182.1, COV2-2196, S2E12, A19-46.1, S309, and LY-CoV1404—that accommodated these changes and neutralized B.1.1.529. We identified combinations of antibodies with synergistic neutralization. The analysis revealed structural mechanisms for maintenance of potent neutralization against emerging variants.

Since first appearing in late 2019 (*1*), severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has infected more than 490 million people and resulted in more than 5.9 million deaths (*2*). The appearance and rapid spread of the B.1.1.529 (Omicron; BA.1) variant (*3*, *4*)—with 34 amino acid substitutions, deletions, and insertions in the spike protein, which is three times higher than found in prior variants—has raised alarm. Although extremely broad antibodies such as S2P6 (*5*) that neutralize diverse β-coronaviruses, including SARS-CoV-2, are likely to be unencumbered by B.1.1.529 mutations, these broad antibodies neutralize in the microgram per milliliter range, whereas current therapeutic antibodies generally neutralize in the 1 to 50 nanogram per milliliter range for the ancestral D614G virus. (Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr. In the mutants, other amino acids were substituted at certain locations; for example, D614G indicates that aspartate at position 614 was replaced by glycine.)

[1]Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, USA. [2]Electron Microscopy Laboratory, Cancer Research Technology Program, Leidos Biomedical Research, Frederick National Laboratory for Cancer Research, Frederick, MD 21702, USA.
*Corresponding author. Email: tzhou@nih.gov (T.Z.); john.misasi@nih.gov (J.M.); pdkwong@nih.gov (P.D.K.)
†These authors contributed equally to this work.
‡These authors contributed equally to this work.

Cryo-EM structure of B.1.1.529 (Omicron) spike

To provide insight into the impact of B.1.1.529 mutations on spike, we expressed and produced the two proline-stabilized (S2P) (*6*) B.1.1.529 spike and collected single-particle cryo–electron microscopy (cryo-EM) data that resulted in a structure of the trimeric ectodomain at 3.29 Å resolution (Fig. 1, fig. S1, and table S1). Like other D614G-containing variants, the most prevalent spike conformation comprised the single–receptor binding domain (RBD)–up conformation (*7*). B.1.1.529 mutations present in the spike gene resulted in three deletions of two, three, and one amino acids, a single insertion of three amino acids, and 30 amino acid substitutions in the spike ectodomain (fig. S2A). As expected from the ~3% variation in sequence, the B.1.1.529 spike structure was extremely similar to the WA-1 spike structure, with an overall Cα-backbone root mean square deviation (RMSD) of 1.8 Å (0.5 Å for the S2 region); however, we did observe minor conformational changes in a few places. For example, the RBD S371L/S373P/S375F substitutions changed the conformation of their residing loop so that F375 in the RBD-up protomer interacted with F486 in the neighboring RBD-down protomer and locked this RBD in down position (Fig. 1B). Moreover, the S373P substitution in the next RBD-down protomer increased contact surface with the neighboring RBD in down position, potentially latching itself in the down position (fig. S1G). All of these S371L/S373P/S375F substitution–mediated interactions help to stabilize the single-RBD-up conformation. Amino acid changes were denser in the N-terminal domain (NTD) and RBD, where most neutralizing antibodies bind, although RMSDs remained low (0.6 and 1.2 Å for NTD and RBD, respectively). About half the B.1.1.529 alterations in sequence outside the NTD and RBD involved new interactions, both hydrophobic, such as Y796 with the glycan on N709, and electrostatic, such as K547 and K856 interacting respectively with residues in heptad repeat 1 (HR1) in S2 and subdomain-1 (SD1) in S1 on neighboring protomers (Fig. 1B, fig. S2A, and table S2). Despite these newly introduced interactions, differential scanning calorimetry indicated that the B.1.1.529 spike had folding energy similar to that of the original WA-1 strain (fig. S2B).

NTD changes altered ~6% of the solvent-accessible surface on this domain, and several were located directly on or proximal to the NTD-supersite of vulnerability (*8*), where prior variants had mutations that substantially reduced neutralization by NTD antibodies. Other NTD changes neighbored a pocket, proposed to be the site of bilirubin binding (*9*), which also binds antibody (Fig. 1C) (*10*).

RBD alterations changed ~16% of the solvent-accessible surface on this domain and were constrained to the outward-facing ridge of the domain (Fig. 1D), covering much of the surface of the trimeric spike apex (fig. S1F). Several amino acid changes involved basic substitutions, resulting in a substantial increase in RBD electropositivity (Fig. 1D). Overall, RBD changes affected binding surfaces for the angiotensin-converting enzyme 2 (ACE2) receptor (Fig. 1D) (*11*) as well as recognition sites for potently neutralizing antibodies (Fig. 1E) (*12*–*14*).

Functional assessment of variant binding to ACE2

When pathogens infect a new species, sustained transmission leads to adaptations that optimize replication, immune avoidance, and transmission. One hypothesis for the efficient species adaptation and transmission of SARS-CoV-2 in humans is that the virus spikes are evolving to optimize binding to the host receptor protein, ACE2. As a first test of this hypothesis, we used a flow cytometric assay to evaluate binding of human ACE2 to cells expressing variant spike proteins. We evaluated the binding of soluble dimeric ACE2 to B.1.1.7 (Alpha) (*15*), B.1.351 (Beta) (*16*), P.1 (Gamma) (*17*, *18*), or B.1.617.2 (Delta) (*19*) spikes compared with the ancestral D614G spike. The early B.1.1.7 variant contains an RBD substitution at N501Y (Fig. 2A), which increases RBD binding to ACE2 (*20*). Consistent with this, cell-surface ACE2 binding to B.1.1.7, which only contains an N501Y substitution in RBD, was 182% of D614G (fig. S3A). However, other N501Y-containing variants (B.1.351 and P.1) and B.1.617.2 (Fig. 2A), which lacks N501Y, did

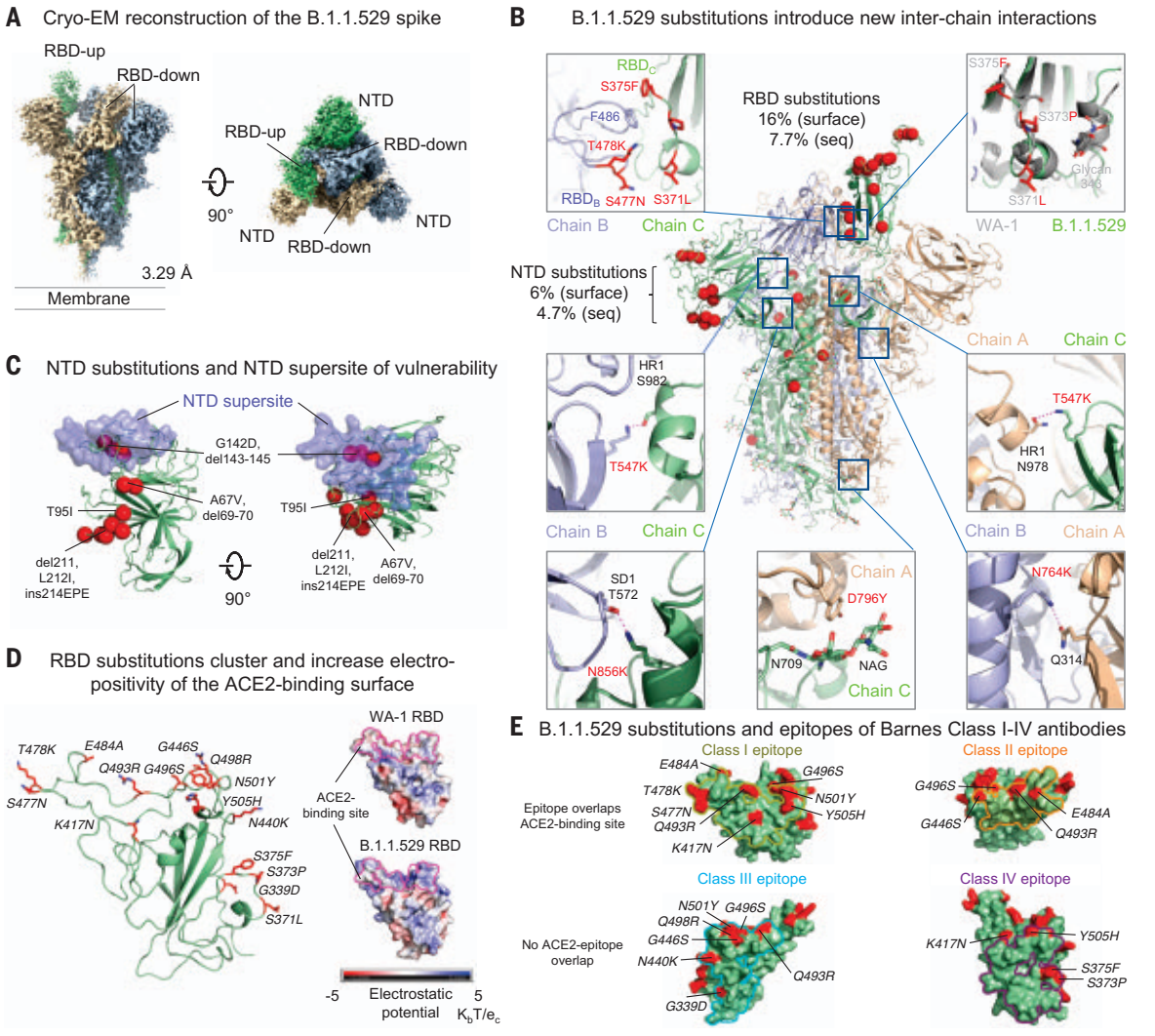


Fig. 1. Cryo-EM structure of the SARS-CoV-2 B.1.1.529 (Omicron) spike. (**A**) Cryo-EM map of the SARS-CoV-2 B.1.1.529 spike. Reconstruction density map at 3.29 Å resolution is shown with side and top views. Protomers are colored light green, wheat, and light blue. The contour level of cryo-EM map is 4.0σ. (**B**) B.1.1.529 amino acid substitutions introduced interprotomer interactions. Substitutions in one of the protomers are shown as red spheres. Examples of interprotomer interactions introduced by B.1.1.529 substitutions are highlighted in the box with zoom-in views to the side. Amino acid substitutions are described as a percentage of the domain surface (surface) or as a percentage of the sequence (seq). (**C**) The NTD supersite of vulnerability is shown in semitransparent surface along with a green backbone ribbon. Amino acid substitutions, deletions, and insertions are in red. (**D**) The 15 amino acid substitutions, clustered on the rim of RBD, changed 16% of the RBD surface area (left) and increased electropositivity of the ACE2-binding site (right). Amino acid substitutions are shown as red sticks. The ACE2-binding site on the electrostatic potential surface are marked as magenta lines. (**E**) Mapping B.1.1.529 RBD substitutions on the epitopes of Barnes class I to IV antibodies. The locations of the substitutions are shown in red on the surface. Those that may potentially affect the activity of antibodies in each class are labeled with their residue numbers. Class I footprint is defined by epitopes of CB6 and B1-182.1; class II footprint is defined by epitopes of A19-46.1 and LY-CoV555; class III footprint is defined by epitopes of A19-61.1, COV2-2130, LY-CoV1404 and S309; and class IV footprint is defined by epitopes of DH1047 and S304. Class I and II antibodies primarily target the ACE2 binding site, whereas the epitopes of class III and IV antibodies do not. Class II and III epitopes allow binding to WA-1 when RBD is in the up or down conformation, although the distinction between class I and II is more fluid, particularly with new variants that alter the accessibility of epitopes relative to WA-1. In addition, some antibodies, such as A19-46.1, can bind fully up intermediate states between up and down but cannot bind the fully down state. We therefore classified primarily by binding region.

not show substantial increases in ACE2 binding signal (fig. S3A), a finding that is consistent with previous reports (*20*). Multiple groups have evaluated ACE2 binding to B.1.1.529 and found both increased or unchanged binding (*21*–*25*). In our cell binding assay, we found that ACE2 binding was 104% of the binding to D614G binding (fig. S3A).

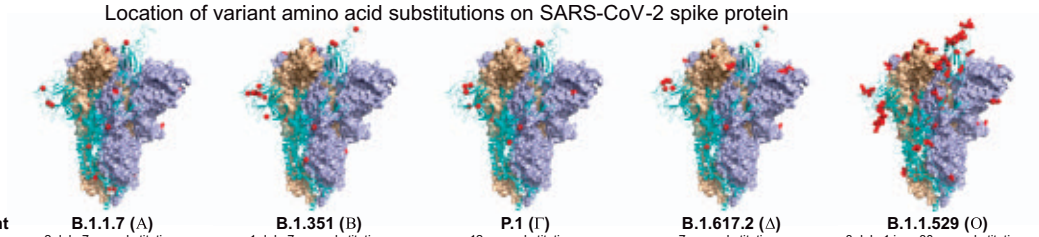
Because cell-surface spike binding may be influenced by factors such as increased electropositivity of the RBD and by relative changes in the up/down state of RBD, we formally investigated the ACE2 binding affinity using surface plasmon resonance of soluble dimeric human ACE2 to S2P spike trimers generated from the ancestral WA-1 and six subsequent variants: D614G, B.1.351, P.1, B.1.617.2, B.1.1.7, and B.1.1.529. We observed that both WA-1 and D614G, which have identical RBD sequences, have similar apparent affinities (K_{app} = 1.1 and 0.73 nM, respectively) (fig. S3, B and C). The apparent affinity for variants was minimally changed (K_{app} = 0.59 to 3.8 nM), including for N501Y-containing variants (fig. S3, B and C). Given the minimal changes to affinity, our data suggests that spike variant evolution is not being driven by the optimization of ACE2 binding but is instead driven primarily by immune pressure.

Variant binding and neutralization by individual monoclonal antibodies

To define the impact of SARS-CoV-2 variant amino acid changes on the binding and neutralization of monoclonal antibodies, we expressed and purified 17 highly potent antibodies targeting the spike RBD (*12*, *13*, *26*–*38*), including 13 antibodies currently under clinical investigation or approved for use under emergency use authorization (EUA) by the US Food and Drug Administration. All antibodies bound and neutralized B.1.1.7 comparable with the ancestral D614G and consistent with the single 501Y substitution being outside each antibody's binding epitope (Fig. 2 and fig. S5A). Consistent with previous reports (*14*, *39*–*42*), two additional RBD substitutions in the RBD of B.1.351 and P.1 variants (Fig. 2A) led to substantially decreased binding and neutralization by the two class I antibodies CB6 and REGN10933 and the two class II antibodies LY-CoV555 and

A

Location of variant amino acid substitutions on SARS-CoV-2 spike protein



Variant	B.1.1.7 (Α)	B.1.351 (Β)	P.1 (Γ)	B.1.617.2 (Δ)	B.1.1.529 (Ο)
Total a.a. substitutions	2 del., 7a.a. substitutions	1 del., 7 a.a. substitutions	12 a.a. substitutions	7 a.a. substitutions	3 del., 1 ins., 30 a.a. substitutions
RBD substitutions	1 RBD substitution N501Y	3 RBD substitutions K417N E484K N501Y	3 RBD substitutions K417T E484K N501Y	2 RBD substitutions L452R T478K	15 RBD substitutions G339D S371L S373P S375F K417N N440K G446S S477N T478K E484A Q493R G496S Q498R N501Y Y505H

B

Cell surface binding of monoclonal antibodies to variant spike proteins

Antibody	Class	D614G	B.1.1.7	B.1.351	P.1	B.1.617.2	B.1.1.529
A23-58.1	I	100	204	116	37	99	13
B1-182.1	I	100	213	118	39	97	9
COV2-2196	I	100	216	121	38	85	3
S2E12	I	100	191	104	35	93	4
CB6	I	100	132	0	0	89	0
REGN10933	I	100	204	11	3	114	0
CT-P59	I	100	197	100	37	105	0
ADG2	I, IV	100	158	98	40	117	24
A19-46.1	II	100	160	102	37	0	47
LY-COV555	II	100	158	0	0	0	0
C144	II	100	141	0	0	90	0
A19-61.1	III	100	146	110	40	107	0
REGN10987	III	100	163	104	42	115	0
COV2-2130	III	100	139	102	34	112	32
C135	III	100	142	106	41	120	0
S309	III	100	139	113	40	122	15
LY-CoV1404	III	100	137	97	36	119	44

Normalized to D614G

200%

100%

0%

C

Pseudotyped virus neutralization of monoclonal antibodies (ng/mL)

			D614G		B.1.1.7		B.1.351		P.1		B.1.617.2		B.1.1.529 (BA.1)	
Antibody	Generic name	Class	IC_{50}	IC_{80}	IC_{50}	IC_{80}	IC_{50}	IC_{80}	IC_{50}	IC_{80}	IC_{50}	IC_{80}	IC_{50}	IC_{80}
A23-58.1	--	I	1.3	4.5	2.1	4.7	4.7	11.6	1.6	5.7	1.6	3.5	231	1132
B1-182.1	--	I	0.9	2.4	1.7	3.9	2.0	4.3	<0.6	1.5	1.0	3.5	281	1301
COV2-2196	tixagevimab *	I	2.0	3.2	2.7	5.4	3.5	10.6	1.9	7.5	1.4	4.5	269	900
S2E12	--	I	1.4	2.9	6.8	3.3	2.2	4.1	<0.6	2.5	1.1	2.4	38.1	112
CB6	etesevimab	I	50.5	109	22.7	141	> 10,000	> 10,000	> 10,000	> 10,000	14.8	492	> 10,000	> 10,000
REGN10933	casirivimab	I	6.1	16.0	9.7	29.1	> 10,000	> 10,000	1536	> 10,000	3.0	3.8	> 10,000	> 10,000
CT-P59	regdanvimab	I	1.5	4.6	5.5	22.3	65.8	233	39.6	153	14.7	78.3	> 10,000	> 10,000
ADG2	--	I,IV	5.1	14.7	4.7	17.1	15.5	43.9	6.4	21.8	7.6	20.8	2037	8113
A19-46.1	--	II	19.4	40.6	39.2	106	57	157	39.7	69.6	> 10,000	> 10,000	223	376
LY-COV555	bamlanivimab	II	3.6	10.4	7.7	24.9	> 10,000	> 10,000	> 10,000	> 10,000	> 10,000	> 10,000	> 10,000	> 10,000
C144	--	II	5.1	9.8	5.7	22.4	> 10,000	> 10,000	> 10,000	> 10,000	4.9	20.6	> 10,000	> 10,000
A19-61.1	--	III	7.7	20.1	31.3	49.4	10.8	19.6	7.3	15.0	18.7	27.5	> 10,000	> 10,000
REGN10987	imdevimab	III	20.0	412	13.5	72	24.4	102	6.5	53.6	455	4926	> 10,000	> 10,000
COV2-2130	cilgavimab *	III	3.7	10.9	6.3	12.5	5.4	13.3	14.1	18.2	25.0	80.5	5850	> 10,000
C135	--	III	10.8	64.3	13.1	191	34.0	n.d. ‡	14.7	115	24.7	n.d. ‡	> 10,000	> 10,000
S309	sotrovimab *	III	36.1	163	30.6	194	27.7	97.4	41.8	359	45.2	113	281	1336
LY-CoV1404	bebtelovimab	III	3.0	5.8	30.6	194	4.1	8.7	11.5	17.6	3.7	8.5	5.1	14.4

Fig. 2. SARS-CoV-2 monoclonal antibody binding and neutralization. (**A**) Models of SARS-CoV-2 WA-1 spike protein (PDB: 6XM3) with the locations of substitutions present in variants indicated as red dots. Also indicated is the total number of amino acid substitutions and the number and locations of RBD substitutions in VOC spike proteins. (**B**) Full-length spike proteins from the indicated SARS-CoV-2 variants were expressed on the surface of transiently transfected 293T cells, and binding to indicated monoclonal antibodies was assessed by means of flow cytometry. Antibody mean fluorescence intensity (MFI) binding signal was adjusted according to spike protein expression level (fig. S4). Shown is the ratio of the adjusted antibody MFI binding to the indicated spike expressing cells to the adjusted MFI of the same antibody bound to D614G spike–expressing cells. The data are expressed as a percentage. Shown is a representative experiment (n = 2 replicates). (**C**) Lentiviruses pseudotyped with SARS-CoV-2 spike proteins from D614G, B.1.1.7, B.1.351, P.1, B.1.617.2, or B.1.1.529 (BA.1) were incubated with serial dilutions of the indicated antibodies, and IC_{50} and IC_{80} values were determined. S309 was tested on 293 flpin-TMPRSS2-ACE2 cells, whereas all the other antibodies were tested on 293T-ACE2 cells. Ranges are indicated with white (>10,000 ng/ml), light blue (>1000 to ≤10,000 ng/ml), yellow (>100 to ≤1000 ng/ml), orange (>50 to ≤100 ng/ml), red (>10 to ≤50 ng/ml), maroon (>1 to ≤10 ng/ml), and purple (≤1 ng/ml). n.d. ‡, not determined because of incomplete neutralization that plateaued at <80% (fig. S5B). Where available, generic names of antibodies under therapeutic investigation are shown. Gray shading indicates antibodies that previously or currently have received Emergency Use Authorization from the US Food and Drug Administration. Generic names with an asterisk indicate therapeutic antibody products with the same binding regions as those of the antibodies being tested but containing amino acid changes in their Fc domains.

C144 (Fig. 2, B and C, and fig. S5A). In addition, whereas binding of CT-P59 to B.1.351 and P.1 variants was minimally changed (37 to 100%), neutralization was decreased 26- to 43-fold (Fig. 2, B and C). The remaining antibodies showed minimal binding changes and a <3.6-fold difference in neutralization half-maximal inhibitory concentration (IC_{50}) (Fig. 2, B and C, and fig. S5A). An evaluation of the antibodies in our panel against B.1.617.2 revealed minimal changes in binding and neutralization for all antibodies except REGN10987, A19-46.1, and LY-CoV555 (Fig. 2, B and C). As previously reported (*14*, *39*–*42*), REGN10987 binds B.1.617.2 spike but has 22-fold less neutralization, and the

binding and neutralizing activity of A19-46.1 and LY-CoV555 was eliminated (Fig. 2, B and C). These data are consistent with previous results that showed both A19-46.1 and LY-CoV555 were sensitive to the L452R mutations present in B.1.617.2 (*14*, *39*, *40*).

For B.1.1.529, all but three antibodies (A19-46.1, COV2-2130, and LY-CoV1404) showed binding less than 32% of D614G. Furthermore, although COV2-2196, S2E12, B1-182.1, and A23-58.1 use the same VH1-58 gene in their heavy chain and target a similar region on the RBD (the VH1-58 supersite), they showed differential binding to B.1.1.529 (3, 4, 9, and 13%, respectively) and B.1.617.2 (85, 93, 97, and 99%, respectively) (Fig. 2B). Even though the absolute differences in binding are minimal, the shared trend may be reflective of how the RBD tip T478K substitution found in B1.1.529 and B1.617.2 is accommodated by each of these antibodies. Taken together, cell surface binding suggests that whereas both A19-46.1 (47%) and LY-CoV1404 (44%) are likely to retain potent neutralizing activity against B.1.1.529, the remaining antibodies in our panel might show decreased neutralizing activity.

Using the same panel of monoclonal antibodies, we further assayed for each antibody's capacity to neutralize the B.1.1.529 variant. VH1-58 supersite antibodies are a subset of class I antibodies that bind to the tip of RBD and have high neutralization activity against previous variants (*14*); despite this, their IC_{50}s were 40- to 126-fold worse against B.1.1.529 relative to D614G (Fig. 2C). In addition, two other antibodies, CB6 (class I) and ADG2 (class I/IV), were severely affected (Fig. 2C). Among the class II antibodies (LY-CoV555, C144, and A19-46.1), neutralization by LY-CoV555 and C144 was completely abolished. By contrast, we found that the A19-46.1 neutralization IC_{50} was 223 ng/ml for B.1.1.529 versus 19.4 ng/ml for D614G (Fig. 2C) and was less than sixfold of the previously reported IC_{50} for WA-1 (39.8 ng/ml) (*14*). For class III antibodies, neutralization activity of A19-61.1, REGN10987, and C135 was completely abolished; CoV2-2130 decreased 1581-fold; and that of S309 decreased by approximately eightfold (Fig. 2C). In contrast to all the other antibodies, we found that the neutralization of LY-CoV1404 against B.1.1.529 was unchanged relative to D614G (Fig. 2C). Taken together, these data demonstrate that the mutations present in B.1.1.529 mediate resistance to a broad range of antibodies.

Structural and functional basis of class I antibody neutralization, escape, and retained potency

To determine the functional basis of B.1.1.529 neutralization and escape for class I antibodies, we analyzed class I antibodies CB6, B1-182.1, and S2E12, which show differential B.1.1.529 neutralization (Fig. 2C). CB6 is a class I antibody that does not use the VH1-58 gene and whose epitope is partially overlapping with the VH1-58 supersite. We used virus particles containing single–amino acid substitutions representing each of 15 single–amino acid changes on the RBD of B.1.1.529. Whereas K417N completely abrogated neutralization of CB6, the presence of Y505H, S371L, or Q493R substitutions decreased neutralization from 7- to 46-fold (Fig. 3A). Taken together, this suggests that B.1.1.529 evades CB6-like antibodies through multiple substitutions. Docking of the RBD-bound CB6 onto the B.1.1.529 structure revealed several B.1.1.529 substitutions that may affect CB6 binding through a steric clash (Q493R) and removal of key contacts (K417N and Y505H), which is consistent with neutralization data (Fig. 3B). The VH1-58 supersite antibodies B1-182.1 and S2E12 have similar amino acid sequences to each other (*14*) but show an approximately sixfold difference in B.1.1.529 neutralization. These two antibodies remained highly potent for all virus particles with single RBD mutations (IC_{50} < 10.6 ng/ml), with the largest change for Q493R, which caused a 7- and 5.4-fold decrease of neutralization for B1-182.1 and S2E12, respectively (Fig. 3A). These small differences in neutralization from single mutations suggest that two or more combinations of mutations of B.1.1.529 are working in concert to mediate escape from VH1-58 supersite antibodies. Docking of the RBD-bound B1-182.1 onto the B.1.1.529 structure indicated that the epitopes of these antibodies were bounded by Q493R, S477N, T478K, and E484A, with R493 pressing on one side of the antibody and N477/K478 on the other side of the antibody at the heavy chain–light chain interface (Fig. 3C). N477/K478 positioned at the junction formed by complementarity-determining region (CDR) H3, CDR L1, and L2 and clashed slightly with a region centered at CDR H3 residue 100C (Kabat numbering) (Fig. 3D). Sequence alignment of CDR H3 of VH1-58–derived antibodies indicated that residue 100C varies in side chain size, from serine in S2E12 to tyrosine in A23-58.1. The size of 100C reversely correlated with neutralization potency IC_{80} (P = 0.046) (Figs. 2C and 3D), suggesting that VH1-58 antibodies could alleviate escape imposed by the B.1.1.529 mutations through reduced side chain size at position 100C to minimize clashes from N477/K478.

Structural and functional basis of class II antibody neutralization, escape, and retained potency

We next sought to determine the functional basis of B.1.1.529 neutralization and escape for two class II antibodies, LY-CoV555 (*31*) and A19-46.1 (*14*), which have B.1.1.529 IC_{50} of >10,000 and 223 ng/ml, respectively (Fig. 2C). Consistent with previous reports (*14*, *43*, *44*), either E484A or Q493R substitution results in complete loss of LY-CoV555 neutralization, whereas the same mutations did not affect A19-46.1 (Fig. 4A). For A19-46.1, no individual mutation reduced neutralization to the level noted in B.1.1.529 except S371L, which increased the IC_{50} to 72.3 ng/ml (Fig. 4A). In the context of B1.1.529, which contains the S371L/S373P/S375F alterations, the IC_{50} further increased to 223 ng/ml (Fig. 2C). One potential explanation for this further reduction of potency is that the mutation-introduced interaction between F375 and F486 (Fig. 1B) restricts the RBD-up conformation required for A19-46.1 binding.

To understand the structural basis of A19-46.1 neutralization of B.1.1.529, we obtained a cryo-EM structure of the B.1.1.529 spike in complex with Fab A19-46.1 at 3.86 Å resolution (Fig. 4B, fig. S6, and table S1). Two Fabs bound to the RBDs in up-conformation in each spike, with the third RBD in down position. Docking Fab A19-46.1 onto the RBD in down conformation revealed a clash with the NTD of the neighboring protomer, suggesting that A19-46.1 binding requires the RBD-up conformation. Focused local refinement of the antibody-RBD region resolved the antibody-RBD interface (Fig. 4B, right). Consistent with previous mapping and negative stain EM data, A19-46.1 binds to a region on RBD generally targeted by class II antibodies with an angle ~45° toward the viral membrane. Binding involves all light chain CDRs and only CDR H3 of the heavy chain and buries a total of 805 $Å^2$ interface area from the antibody (Fig. 4C, left). With the light chain latching to the outer rim of the RBD and providing about 70% of the binding surface, A19-46.1 uses its 17-residue-long CDR H3 to form parallel strand interactions with RBD residues 345 to 350 (Fig. 4B, right). Docking RBD-bound ACE2 to the A19-46.1–RBD complex indicated that the bound antibody sterically clashes with ACE2 (Fig. 4D), providing the structural basis for its neutralization of B.1.1.529.

The 686 $Å^2$ epitope of A19-46.1 is located within an RBD region that is not mutated in B.1.1.529. Three of the 15 amino acid changes on the RBD, S446, A484, and R493 are positioned at the edge of epitope, with their side chains contributing 8% of the binding surface. LY-CoV555, which targets the same region as that of a class II antibody, completely lost activity against B.1.1529. Superimposing the LY-CoV555-RBD complex onto the B.1.1.529 RBD showed that although LY-CoV555 has an angle of approach similar to that of A19-46.1 (Fig. 4D), its epitope shifts up to the ridge of the RBD and includes the B.1.1.529 alterations A484 and R493 (Fig. 4D). R493 causes steric clash with the CDR H3 of LY-CoV555, explaining the escape of B.1.1.529 from LY-CoV555 neutralization. Overall, the location

A Pseudotyped virus neutralization by B1.182.1, S2E12 and CB6 of B.1.1.529 with single-residue substitutions in RBD (ng/mL)

		D614G	G339D	R346K	S371L	S373P	S375F	K417N	N440K	G446S	S477N	T478K	E484A	Q493R	G496S	Q498R	N501Y	Y505H
IC_{50}	B1-182.1	1.5	2.9	1.8	5.3	2.5	1.2	0.8	1.5	1.1	2.7	1.4	3.7	10.6	1.9	1.6	2.0	2.5
	S2E12	1.0	2.4	1.2	2.9	2.0	1.8	1.0	1.2	1.6	2.8	1.3	2.5	5.4	2.4	1.4	2.0	2.8
	CB6	6.9	12.3	16.8	212	13.6	36.6	>10,000	5.9	10.8	6.8	5.1	26.6	320	24	32	22	50
IC_{80}	B1-182.1	4.9	5.3	4.7	20.0	4.7	2.8	2.4	3.9	3.8	5.1	3.6	14.0	23.2	5.8	3.7	3.7	4.9
	S2E12	3.2	4.8	2.6	6.4	3.8	4.3	2.6	4.9	2.9	4.9	2.4	6.3	17.1	4.4	3.2	3.7	6.2
	CB6	37.8	59.1	55.1	747	55.7	199	>10,000	60.2	45.3	41.1	33.8	124	1315	92	125	179	167

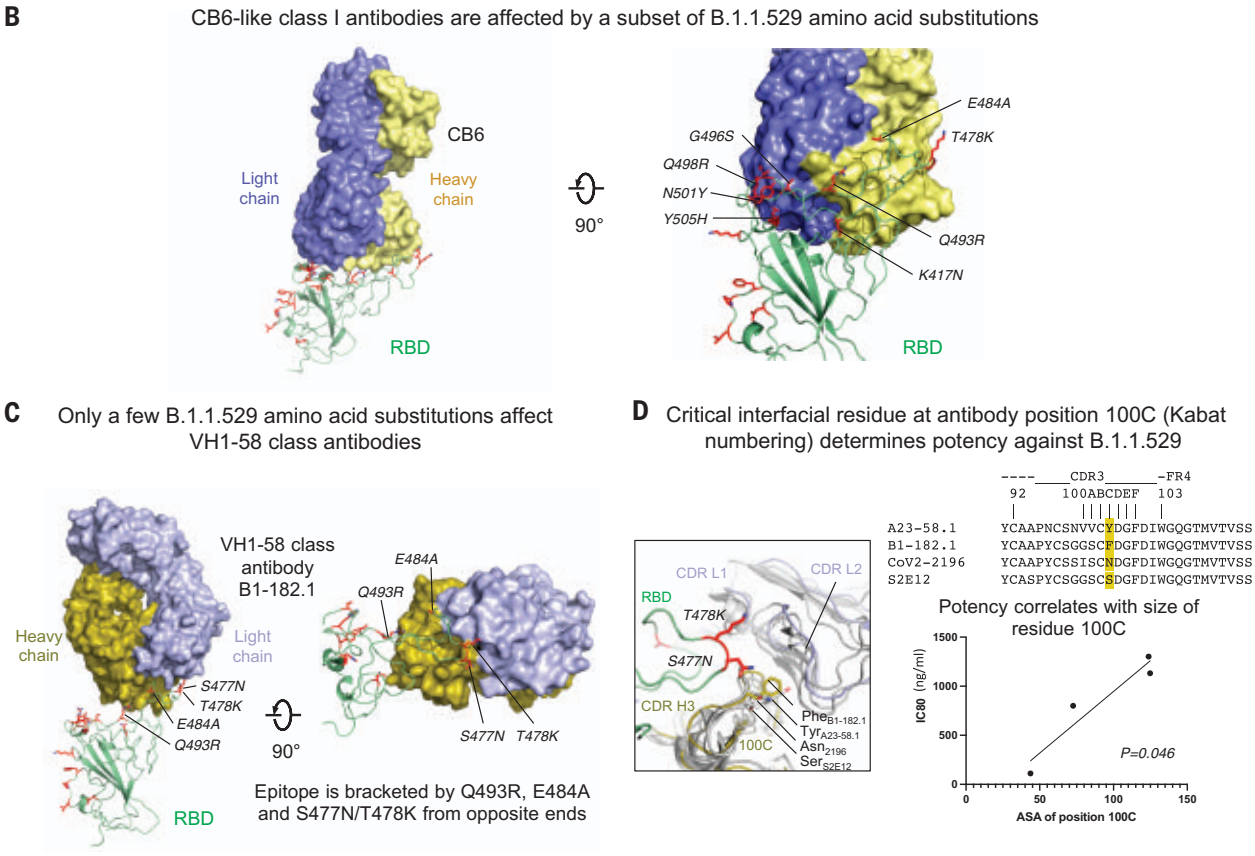


Fig. 3. Functional and structural basis of class I antibody neutralization and mechanistic basis of retained potency against B.1.1.529 VOC. (**A**) Lentiviruses pseudotyped with SARS-CoV-2 spike proteins from D614G or D614G plus the indicated point substitutions found within the B.1.1.529 spike were incubated with serial dilutions of the indicated antibodies, and IC_{50} and IC_{80} values were determined on 293T-ACE2 cells. Ranges are indicated with white (>10,000 ng/ml), light blue (>1000 to ≤10,000 ng/ml), yellow (>100 to ≤1000 ng/ml), orange (>50 to ≤100 ng/ml), red (>10 to ≤50 ng/ml), maroon (>1 to ≤10 ng/ml), and purple (≤1 ng/ml). (**B**) Mapping of B.1.1.529 amino acid substitutions at the epitope of class I antibody CB6. RBD-bound CB6 was docked onto the B.1.1.529 spike structure. B.1.1.529 amino acid substitutions incompatible with CB6 binding were identified and labeled. The K417N substitution caused a clash in the center of the paratope. B.1.1.529 RBD is shown in green cartoon, with amino acid substitutions as red sticks. CB6 is shown in surface representation, with heavy and light chains in yellow and slate, respectively. (**C**) Docking of RBD-bound VH1-58–derived class I antibody B1-182.1 onto the B.1.1.529 spike structure identified four substitutions with potential steric hindrance. B1-182 is shown in surface representation, with heavy and light chains colored olive and light blue, respectively. B.1.1.529 amino acid substitutions that may affect binding of VH1-58 antibodies were labeled. (**D**) Structural basis for effective neutralization of the B.1.1.529 VOC by VH1-58–derived antibodies. Even though VH1-58 antibodies—such as the S2E12, COV2-2196, A23-58.1, and B1-182.1—share high-sequence homology (top right), their neutralization potency against B.1.1.529 varies. Structural analysis indicated that CDR H3 residue 100C, located at the interface formed between RBD and antibody heavy and light chains, may determine their potency against B.1.1.529 (left). Size of this residue correlated with neutralization potency with two-tailed $P = 0.046$ (bottom right).

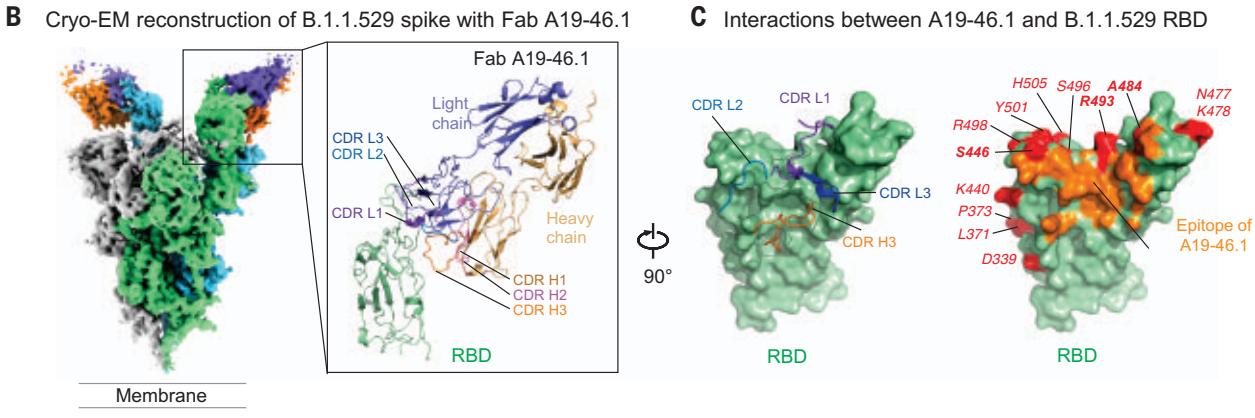
of the epitope and the angle of approach allow A19-46.1 to effectively neutralize B.1.1.529.

Structural and functional basis of class III antibody neutralization, escape, and retained potency

To evaluate the functional basis of B.1.1.529 neutralization and escape for class III antibodies and to understand how potent neutralization might be retained, we investigated a panel of class III antibodies with differential potency, including A19-61.1, COV2-2130, S309, and LY-CoV1404 (Fig. 5A). Assessment of the impact of each of the 15 mutations in the RBD revealed that the G446S amino acid change results in a complete loss in activity for A19-61.1 (Fig. 5A), which is consistent with the complete loss of function of this antibody against B.1.1.529 and previous reports that suggested G446V might affect function (*14*). For S309, S373P resulted in a small change in neutralization (Fig. 5A). Unexpectedly, although S309 retains moderate neutralizing activity against B.1.1.529, we found that the single S371L amino acid change leads to a loss in S309 neutralization (Fig. 5A). This suggests that combinations of S371L with other B.1.1.529

A Pseudotyped virus neutralization by A19-46.1 and LY-CoV555 of B.1.1.529 with single-residue substitutions in RBD (ng/mL)

		D614G	G339D	R346K	S371L	S373P	S375F	K417N	N440K	G446S	S477N	T478K	E484A	Q493R	G496S	Q498R	N501Y	Y505H
IC_{50}	LY-COV555	7.3	10.4	17.7	9.3	7.0	3.5	3.5	9.3	7.3	10.3	14.1	>10,000	>10,000	1.8	16.3	15.1	7.5
	A19-46.1	17.8	20.0	17.7	72.3	19.0	10.8	10.8	18.4	30.8	17.3	24.2	16.6	24.8	15.6	40.5	17.1	16.0
IC_{80}	LY-COV555	16.0	21.8	44.2	22.4	14.3	9.7	9.7	23.4	17.8	20.5	40.0	>10,000	>10,000	8.8	42.7	34.7	15.5
	A19-46.1	41.7	47.4	44.2	159	62.7	25.3	25.3	37.1	60.9	37.5	66.9	34.9	62.0	37.6	67.2	38.4	44.6



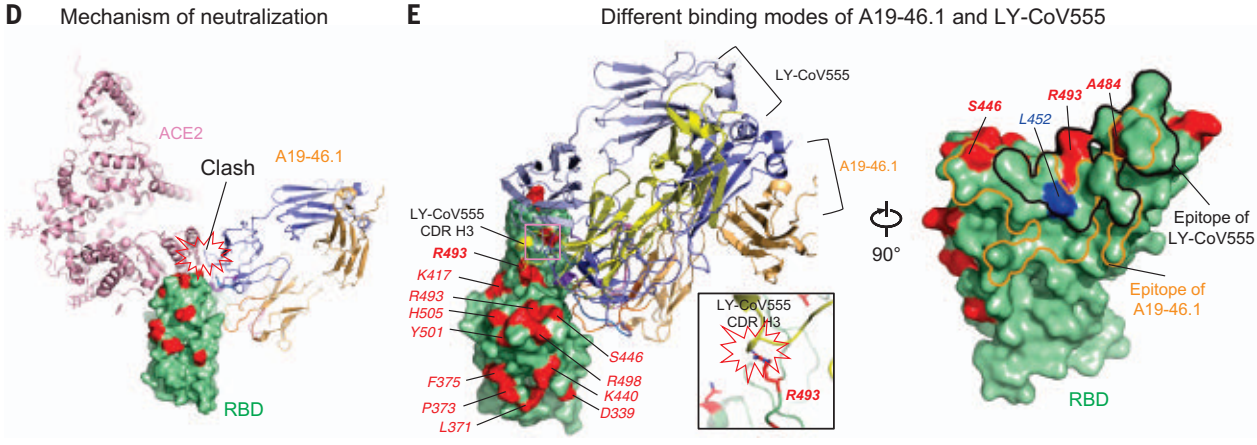


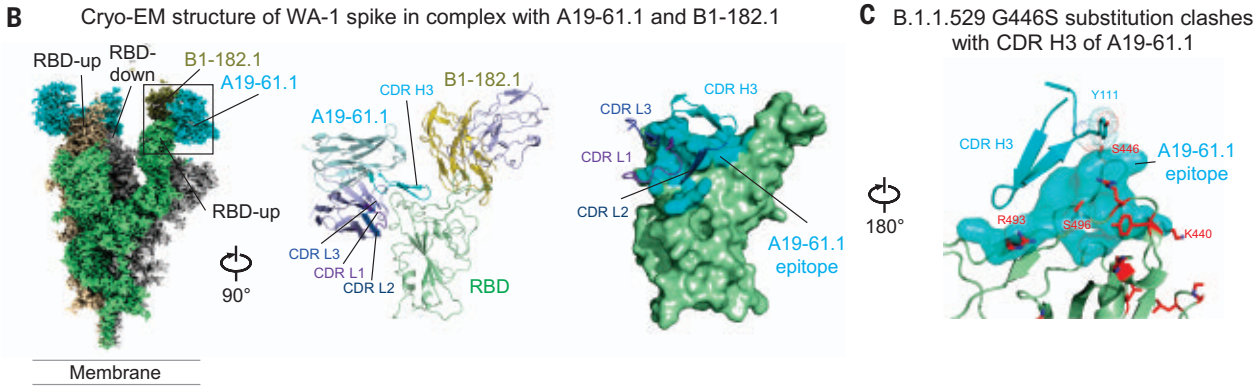
Fig. 4. Functional and structural basis of class II antibody binding, neutralization, and escape. (**A**) Lentiviruses pseudotyped with SARS-CoV-2 spike proteins from D614G or D614G plus the indicated point substitutions found within the B.1.1.529 spike were incubated with serial dilutions of the indicated antibodies, and IC_{50} and IC_{80} values were determined on 293T-ACE2 cells. Ranges are indicated with white (>10,000 ng/ml), light blue (>1000 to ≤10,000 ng/ml), yellow (>100 to ≤1000 ng/ml), orange (>50 to ≤100 ng/ml), red (>10 to ≤50 ng/ml), maroon (>1 to ≤10 ng/ml), and purple (≤1 ng/ml). (**B**) Cryo-EM structure of class II antibody A19-46.1 Fab in complex with the B.1.1.529 spike. Overall density map is shown to the left, with protomers in light green, gray, and light cyan. Two A19-46.1 Fabs bound to the RBD in the up conformation are shown in orange and slate. Structure of the RBD and A19-46.1 after local focused refinement is shown to the right in cartoon representation. The heavy-chain CDRs are in brown, pink, and orange for CDR H1, CDR H2, and CDR H3, respectively. The light chain CDRs are in marine purple blue, marine blue, and blue for CDR L1, CDR L2, and CDR L3, respectively. The contour level of the cryo-EM map is 4.0σ. (**C**) Interaction between A19-46.1 and RBD. (Left) CDR H3 and all light-chain CDRs that are involved in binding of RBD. Epitope of A19-46.1 is shown in orange on the green B.1.1.529 RBD surface, with amino acid substitutions in red. (Right) S446, A484, and R493 are located at the edge of the epitope of Fab A19-46.1. RBD residues are labeled with italicized font. (**D**) Binding of A19-46.1 to RBD prevents binding of the ACE2 receptor. ACE2 and A19-46.1 are shown in cartoon representation. (**E**) Comparison of binding modes to RBD for antibody A19-46.1 and LY-CoV555. (Left and inset) Even though both antibodies target similar regions on RBD, different approaching angles caused a clash between LY-CoV555 CDR H3 and B.1.1.529 substitution R493. (Right) B.1.1.529 substitutions involved in binding of A19-46.1 are only at the edge of its epitope, whereas both R493 and A484 locate in the middle of LY-CoV555 epitope. L452R substitution that eliminates A19-46.1 and LY-CoV555 binding in other SARS-CoV-2 variants is in blue.

mutations result in structural changes in spike that allow S309 to partially overcome the S371L change. None of the single–amino acid changes evaluated resulted in markedly different neutralization by COV2-2130 (Fig. 5A), suggesting that combinations of amino acid substitutions act in concert to decrease neutralization potency against B.1.1.529. Last, consistent with the overall high potency of LY-CoV1404 against all tested variants of concern (VOCs), we did not identify an amino acid change that affected its function.

To understand the structural basis of class III antibody neutralization and viral escape, we determined the cryo-EM structure of WA-1 S2P in complex with Fab A19-61.1 (and Fab B1-182.1 to aid EM resolution of local refinement) at 2.83 Å resolution (Fig. 5B, fig. S8, and table S1). The structure revealed that two RBDs were in the up-conformation with both antibodies bound, and the third RBD was in the down-position with only A19-61.1 bound, indicating that A19-61.1 could recognize RBD in both up and down conformation (Fig. 5B). Local refinement of the RBD-Fab A19-61.1 region showed that A19-61.1 targets the class III epitope with interactions provided by the 18-residue-long CDR H3 from the heavy chain and all CDRs from the light chain (Fig. 5B). Docking the A19-61.1 structure to the B.1.1.529 spike structure

A Pseudotyped virus neutralization by A19-61.1, S309, CoV2-2130 and LY-CoV1404 of B.1.1.529 with single-residue substitutions in RBD (ng/mL)

		D614G	G339D	R346K	S371L	S373P	S375F	K417N	N440K	G446S	S477N	T478K	E484A	Q493R	G496S	Q498R	N501Y	Y505H
IC_{50}	A19-61.1	9.8	18.5	14.7	7.4	8.2	11.7	10.0	14.2	>10,000	16.1	7.7	15.6	13.7	7.0	15.5	11.4	14.5
	S309	41.7	52.1	67.3	>10,000	103	59.2	23.9	82.8	67.7	36.5	38.9	25.8	57.7	29.1	36.8	76.1	38.5
	COV2-2130	2.6	3.6	5.2	2.4	2.6	1.7	1.1	3.3	4.7	2.3	2.9	3.1	1.7	1.6	5.9	3.3	4.7
	LY-CoV1404	3.1	5.1	3.4	1.4	3.3	1.1	3.3	2.6	4.1	4.7	2.6	3.5	3.1	2.7	2.5	3.1	4.4
IC_{80}	A19-61.1	21.3	35.6	16.3	16.8	19.9	40.6	22.0	21.1	>10,000	17.6	16.5	17.1	20.4	20.7	17.0	21.7	26.8
	S309	888	208	631	>10,000	483	365	245	758	424	631	294	228	554	92.6	444	655	131
	COV2-2130	4.6	7.6	14.5	9.1	5.4	6.7	3.4	7.7	16.2	5.2	5.3	7.0	4.8	8.1	14.5	7.7	9.9
	LY-CoV1404	6.2	12.8	7.4	3.8	5.7	3.7	8.4	5.6	10.6	9.9	6.1	7.9	7.0	7.9	4.4	5.3	7.8



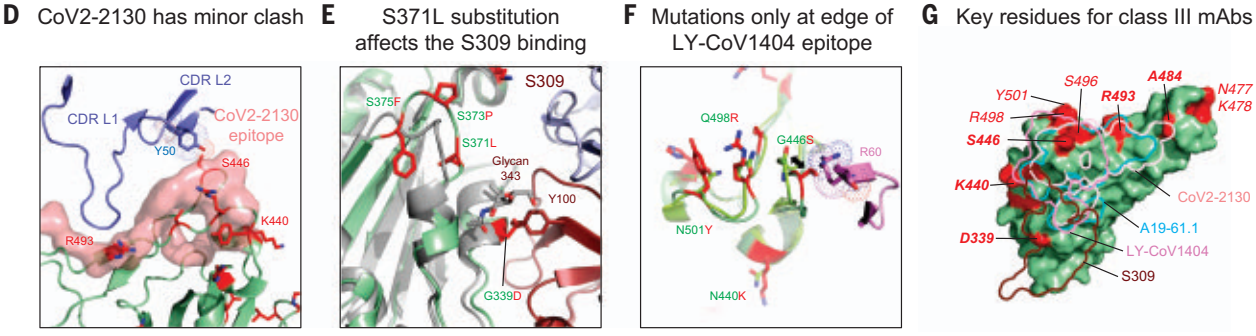


Fig. 5. Functional and structural basis of class III antibody binding, neutralization, and retained potency against the B.1.1.529 VOC. (**A**) Lentiviruses pseudotyped with SARS-CoV-2 spike proteins from D614G or D614G plus the indicated point substitutions found within the B.1.1.529 spike were incubated with serial dilutions of the indicated antibodies, and IC_{50} and IC_{80} values were determined. A19-61.1 and LY-COV1404 were assayed on 293T-ACE2 cells, whereas S309 and CoV2-2130 were tested on 293 flpin-TMPRSS2-ACE2 cells. Ranges are indicated with white (>10,000 ng/ml), light blue (>1000 to ≤10,000 ng/ml), yellow (>100 to ≤1000 ng/ml), orange (>50 to ≤100 ng/ml), red (>10 to ≤50 ng/ml), maroon (>1 to ≤10 ng/ml), and purple (≤1 ng/ml). (**B**) Cryo-EM structure of SARS-CoV-2 WA-1 spike in complex with class I antibody B1-182.1 and class III antibody A19-61.1 at 2.83 Å resolution. Overall density map is shown, with protomers in light green, gray, and wheat. Two RBDs are in the up conformation, with each binding both Fabs, and one RBD is in the down position, with A19-61.1 bound. (Left) RBD. B1-182.1 and A19-61.1 are in olive and cyan, respectively. Structure of the RBD with both Fabs bound after local focused refinement is shown to the right in cartoon representation. (Middle) RBD is shown in green cartoon, and antibody light chains are in light blue. (Right) Epitope of A19-61.1 is shown as cyan surface on RBD, with interacting CDRs labeled. The contour level of cryo-EM map is 5.2σ. (**C**) Structural basis of B.1.1.529 resistance to A19-61.1. Mapping of the A19-61.1 epitope onto the B.1.1.529 RBD indicated that G446S clashed with CDR H3 of A19-61.1. RBD is shown in green cartoon, with amino acid substitutions as red sticks, and epitope of A19-61.1 is the cyan surface. (**D**) Structural basis of CoV2-2130 neutralization of the B.1.1.529 VOC. Docking of the CoV2-2130 onto the B.1.1.529 RBD showed that Y50 in CDR L2 posed a minor clash with S446. RBD is shown in green cartoon, with amino acid substitutions as red sticks, and epitope of CoV2-2130 is the pink surface. (**E**) Structural basis of S309 neutralization of the B.1.1.529 VOC. Docked complex of S309 and B.1.1.529 RBD showed that the S371L/S373P/S375F Loop changed conformation, and the S371L substitution is adjacent to the S309 epitope, whereas the G339D substitution is located inside the epitope. D339 side-chain clashes with CDR H3 Y100. B.1.1.529 RBD is shown in green cartoon, with amino acid substitutions as red sticks, and WA-1 RBD is shown in gray cartoon. (**F**) Structural basis of LY-CoV1404 neutralization of the B.1.1.529 VOC. Docking of the LY-CoV1404 onto the B.1.1.529 RBD identified four amino acid substitutions in the epitope, with G446S causing a potential clash with CDR H2 R60. However, comparison of both LY-CoV1404–bound and –nonbound B.1.1.529 RBD indicated that the S446 loop has the flexibility to allow LY-CoV1404 binding. B.1.1.529 residues at LY-CoV1404 epitope are shown as red sticks, with corresponding WA-1 residues as green sticks. CDR H3 is shown in cartoon representation and colored magenta. (**G**) Overlay of epitope footprints of class III antibodies onto the B.1.1.529 RBD. Locations of amino acid substitutions in B.1.1.529 RBD are in red on the green surface.

indicated that B.1.1.529 mutations S446, R493, and S496 might interfere with A19-61.1. Analysis of the side-chain interactions identified a clash between Y111 in CDR H3 and S446 in the RBD that could not be resolved with loop flexibility (Fig. 5C), explaining the loss of A19-61.1 neutralization against G446S-containing SARS-CoV-2 variants.

Neutralization assays indicated that among the class III antibodies, COV2-2130, S309, and LY-CoV1404 showed variable neutralization potency against B.1.1.529. Docking indicated that CoV2-2130 targets an epitope very similar to A19-61.1, with interactions mainly mediated by its CDR L1 and L2 and avoiding close contact

with R493 and S496. However, the OH group of Y50 in CDR L2 showed a minor clash with S446 in RBD, which explains the structural basis for the partial conservation of neutralization by CoV2-2130 (Fig. 5D). Antibody S309 showed higher potency against B.1.1.529 than CoV2-2130. In a docked complex, the G339D mutation is located inside the epitope and clashes with CDR H3 Y100; however, the void space between S309 and RBD might accommodate an alternate tyrosine rotamer. The S371L/S373P/S375F mutations changed the conformation of their residing loop and may push the glycan on N343 toward S309 to reduce binding (Fig. 5E). LY-CoV1404 was not affected by B.1.1.529 mutations. Docking of the LY-CoV1404 onto the B.1.1.529 RBD identified four amino acid substitutions located at the edge of its epitope. Three of the residues—K440, R498, and Y501—only make limited side-chain interactions with LY-CoV1404. The fourth residue, G446S, caused a potential clash with CDR H2 R60. However, comparison of LY-CoV1404–bound and –nonbound RBD indicated that the loop containing S446 had conformational flexibility that could allow LY-CoV1404 binding (Fig. 5F). Overall, the epitopes to class III antibodies were mainly located on mutation-free RDB surfaces with edges contacting a few B.1.1.529 alterations (Fig. 5G). LY-CoV1404 retained high potency by accommodating all four B.1.1.529 alterations at the edge of its epitope by exploiting loop mobility or by minimizing side-chain interactions.

Synergistic neutralization by the combination of B1-182-1 and A19-46.1

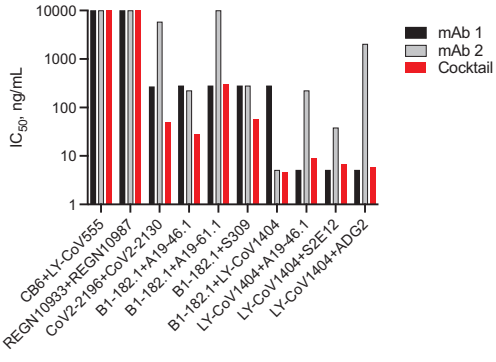
We previously reported that the combination of B1-182.1 and either A19-46.1 or A19-61.1 mitigated mutational escape in an in vitro virus escape assay (*14*), which suggests the possibility of synergistic neutralization. To look for other synergistic combinations, we determined the neutralization of B.1.1.529 pseudotyped viruses through clinically used cocktails or various combinations of B1-182.1, A19-46.1, A19-61.1, LY-CoV1404, ADG2, and S309. Of the 10 combinations evaluated, only COV2-2196+COV2-2130, B1-182.1+A19-46.1, and B1-182.1+S309 neutralized B.1.1.529 with an appreciably improved potency (IC_{50} of 50.8, 28.3, and 58.1 ng/ml, respectively) over the individual component antibodies (Fig. 6, A and B). Each of these included a VH-158 supersite antibody and showed a five- to 115-fold improvement over the component antibodies (Fig. 6B), suggesting an effect that is more than an additive for the specific combination against B.1.1.529.

To understand the structural basis of the improved neutralization by the cocktail of B1-182.1 and A19-46.1, we determined the cryo-EM structure of the B.1.1.529 S2P spike in complex with Fabs of B1-182.1 and A19-46.1 at 3.86 Å resolution (Fig. 6C, fig. S9, and table S1). Three-dimensional reconstruction revealed that the combination of these two antibodies induced the spike to a three-RBD-up conformation, with both Fabs bound to each RBD (although Fabs on one of the RBDs were lower in occupancy). The spike had a 1.6 Å RMSD relative to the three-RBD-up WA-1 structure [Protein Data Bank (PDB) ID: 7KMS]. Overall, the structure showed that these two antibodies were capable of simultaneously recognizing the same RBD, and the combination increased the overall stoichiometry compared with two Fabs per trimer observed in the S2P-A19-46.1 structure described above. Of all the antibodies tested, all VH1-58–derived antibodies retained reasonable levels of neutralization against B.1.1.529, whereas some members of other antibody classes suffered complete loss of activity. VH1-58 antibodies have few alterations in

A Potent neutralization by antibody cocktails

	Class	B.1.1.529 IC_{50}	B.1.1.529 IC_{80}
CB6+LY-CoV555	I/II	> 10,000	> 10,000
REGN10933+REGN10987	I/III	> 10,000	> 10,000
COV2-2196+COV2-2130	I/III	50.8	131
B1-182.1+A19-46.1	I/II	28.3	65.7
B1-182.1+A19-61.1	I/III	311	426
B1-182.1+S309	I/III	58.1	194
B1-182.1+LY-CoV1404	I/III	4.6	14.4
LY-CoV1404+A19-46.1	III/II	8.9	21.7
LY-CoV1404+S2E12	III/I	6.8	23.7
LY-CoV1404+ADG2	III/I, IV	5.8	19.2

B Evaluation of improvement in antibody neutralization



C Cryo-EM structure of B.1.1.529 spike in complex with A19-46.1 and B1-182.1 reveal a 2-RBD-up conformation, with both antibodies binding to both 'up' RBDs

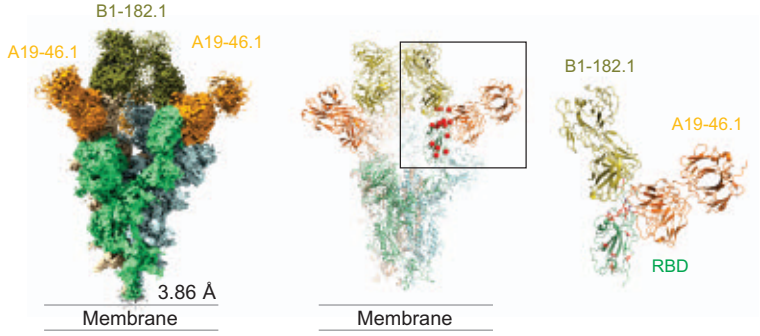


Fig. 6. Potent neutralization of SARS-CoV-2 B.1.1.529 using combinations of antibodies. (**A**) Lentiviruses pseudotyped with SARS-CoV-2 B.1.1.529 spike were incubated with serial dilutions of the indicated combination of antibodies, and IC_{50} and IC_{80} values were determined. Ranges are indicated with white (>10,000 ng/ml), light blue (>1000 to ≤10,000 ng/ml), yellow (>100 to ≤1000 ng/ml), orange (>50 to ≤100 ng/ml), red (>10 to ≤50 ng/ml), maroon (>1 to ≤10 ng/ml), and purple (≤1 ng/ml). (**B**) Neutralization IC_{50} (ng/ml) values for each of the indicated cocktail (*x* axis) or its component antibodies. The IC_{50} for first antibody is listed as mAb1 (black), and the second antibody is listed as mAb2 (gray) or cocktail (red). (**C**) Cryo-EM structure of B.1.1.529 spike in complex with antibodies A19-46.1 and B-182.1 at 3.86 Å resolution. (Left) Overall density map, with protomers colored light green, wheat, and light cyan. (Middle) All RBD are in up conformation, with both Fabs bound. (Right) Binding of one Fab (such as B1-182.1) induces RBD into the up conformation and potentially facilitates binding of the other Fab (such as A19-46.1), which only recognizes the up conformation of RBD. A19-46.1 and B-182.1 are in orange and olive, respectively. The contour level of cryo-EM map is 6.5σ.

their B1.1.529 epitopes and can evolve means to alleviate the impact. We propose a model for the B1-182.1 and A19-46.1 cocktail, in which binding of the first antibody induces the spike into an RBD-up conformation and thereby facilitates the binding of the second antibody that prefers the up conformation. This would kinetically favor an all–RBD-up state by "trapping" RBD in the up position and could lead to a synergistic increase in neutralization potency compared with that of the individual antibodies. This model of synergistic neutralization of Omicron-related viruses and other combinations of antibodies that target the RBD will need further investigation.

Discussion

SARS-CoV2 VOCs provide a window into the coevolution of key host-pathogen interactions between the viral spike, human ACE2 receptor, and humoral immune responses. The RBD is a major target for neutralizing antibodies in both convalescents and vaccinees. An understanding of how RBD mutations evolve may guide the development and maintenance of effective antibody therapeutics and vaccines.

We found that in the context of trimeric spike proteins, variant amino acid changes did not provide a biologically meaningful alteration in affinity to ACE2. When binding trimeric spike protein to immobilized ACE2, our analysis showed that the apparent affinity of B.1.1.529 to ACE2 only changed approximately threefold compared with that of WA-1 (K_{Dapp} = 3.8 nM versus 1.1 nM for WA-1), which is consistent with the 1.4-fold observed by Mannar *et al.* (2.1 nM versus 3.0 nM) (*21*). When tested in the context of RBD, affinity to immobilized ACE2 also showed less than twofold variation between B.1.1.529 and WA-1 (*22*–*25*). This suggests that there is either no further fitness benefit to be gained by improving affinity, that affinity improving changes are being used to compensate for mutations that are deleterious for ACE2 binding but allow immune escape, or both.

Our findings for the class I VH1-58 supersite showed that B.1.1.529 acquires a series of mutations that are not individually deleterious yet bracket the antibody and reduce its potency. VH1-58 antibodies can alleviate the impact by reducing the size of CDR H3 residue 100C to avoid clashes from B.1.1.529 mutations. Because VH1-58 supersite are among the most potent and broadly neutralizing antibodies to SARS-CoV-2 (*14*, *30*, *34*, *45*), our findings point the way toward structure-based designs of existing antibodies to mitigate against amino acid changes at these positions.

For the class II antibody A19-46.1, its preference to RBD in the up-conformation is different from LY-CoV555, which recognizes both RBD-up and -down conformation. The angle of approach and a long-CDR H3 allow A19-46.1 to target the mutation-free face on RBD and minimize contact with mutations on the RBD ridge of B.1.1.529. Comparing the effect of S371L on neutralization by A19-46.1 and LY-CoV555 (Fig. 4A) suggested that L371 (and potentially P373/F375) is critical for controlling the RBD-up or -down conformation in B.1.1.529. This concept is supported by the finding that combination with a class I antibody (such as B1-182.1) synergistically enhances A19-46.1 neutralization (Fig. 6A).

For class III antibodies, only one prototype antibody showed complete loss of B.1.1.529 neutralization. We determined that viral escape was mediated by the G446S amino acid change. This result indicates that potent class III antibodies might be induced through structure-based vaccine designs that mask residue 446 in RBD. Additionally, the existence of G446S-sensitive and -resistant antibodies with substantial epitope overlap suggest that spikes with the G446S substitution can be used to evaluate the quality of class III immune response in serum-based epitope mapping assays (*46*, *47*). In addition, we found that although S309 is severely affected by the S371L mutation alone, it is rescued by compensating mutations in B.1.1.529. Similar but less severe results were recently reported for Brii-198 (*48*, *49*). Taken together, this suggests that there may be a fitness advantage to the virus to maintain a surface that is compatible with S309 binding.

Our analysis of antibodies of clinical importance is consistent with previous reports (*37*, *50*–*52*) and showed that S309 and COV2-2196 neutralized B.1.1.529 to similar degrees. We report that unlike other antibodies, the highly potent LY-CoV1404 does not lose neutralization potency against B.1.1.529. In addition, most antibodies in our panel neutralized the recently described BA.2 Omicron variant (*53*) with similar potency (fig. S10, A and B). The exceptions were with the class III antibodies A19-61.1 and COV2-2130—which fully recovered their neutralization potency, potentially because of the absence of the G446S mutation in BA.2—and S309, which lost more than fivefold activity (fig. S10B).

We identified combinations of antibodies that show more than additive increases in neutralization against B.1.1.529—including COV2-2196+COV2-2130, B1-182.1+A19-46.1, and B1-182.1+S309—and all but the B1-182.1+S309 also show synergy against BA.2 (fig. S10C). Each pair contains a VH1-58 supersite antibody that only binds RBD in the up position and have been shown to be able to bind to all three RBD-up protomers (*14*). We speculate that antibodies that are not affected by S371L, such as VH1-58 mAbs, induce and stabilize the three–RBD-up conformation. This allows antibodies that prefer RBD-up conformation—and would otherwise be unable to break the RBD-down locking conformation imposed by the mutations at 371, 373, and 375—to more efficiently bind. This identification of SARS-CoV-2 monoclonal antibodies that function cooperatively is similar to that seen previously for other viruses (*54*) and supports the concept of using combinations to both enhance potency and mitigate the risk of escape.

Materials and methods

Expression and purification of proteins

Soluble 2P-stabilized SARS-CoV-2 spike proteins were expressed by transient transfection (*6*, *55*). Briefly, plasmid was transfected using Expifectamine (Gibco, #A14525) into Expi293F cells (Gibco, #A14527) and the cultures enhanced 16 to 24 hours post-transfection. Following 4 to 5 days incubations at 120 rotations per minute (rpm), 37°C, 9% CO2, supernatant was harvested, clarified via centrifugation, and buffer exchanged into 1X PBS. Protein of interests were then isolated by affinity chromatography using Ni-NTA resin (Roche, #589380101) followed by size exclusion chromatography on a Superose 6 increase 10/300 column (GE healthcare, #29091596)).

Expression and purification of biotinylated S2P used in binding studies were produced by an in-column biotinylation method as previously described (*55*). Using full-length SARS-Cov2 S and human ACE2 cDNA ORF clone vector (Sino Biological, #HG10108-CH) as the template to the ACE2 dimer proteins. The ACE2 PCR fragment (1~740aa) was digested with Xbal (New England Biolabs, #R3136S) and BamHI (New England Biolabs, #R0145S) and cloned into the VRC8400 with Avi-HRV3C-single chain-human Fc-his (6x) tag on the C-terminal. All constructs were confirmed by sequencing. Proteins were expressed in Expi293 cells by transfection with expression vectors encoding corresponding genes. The transfected cells were cultured in shaker incubator at 120 rpm, 37°C, 9% CO2 for 4 to ~5 days. Culture supernatants were harvested and filtered, and proteins were purified through a Hispur Ni-NTA resin (Thermo Scientific, #88221) and following a Hiload 16/600 Superdex 200 column (GE healthcare, #28989335) according to manufacturer's instructions. The protein purity was confirmed by means of SDS–polyacrylamide gel electrophoresis (SDS-PAGE).

Synthesis, cloning and expression of monoclonal antibodies

A19-46.1, A19-61.1, B1-182.1, and A23-58.1 were synthesized, cloned and expressed as an immunoglobulin G1 (IgG1) containing an HRV3C protease site as previously reported (*14*). For all other antibodies, variable lambda and kappa light chain sequences were human codon optimized, synthesized and cloned into CMV/R-based lambda or kappa chain expression vectors, as appropriate (Genscript). ADG2 was kindly provided by Dr. Laura M Walker

(Adagio Therapeutics, Waltham, MA) (*28*) and LY-CoV1404 by Dr. Stefanie Žentelis, Dr Emilie Lameignere and Kathryn Westendorf, MSc (AbCellera, Canada) (*29*). Previously published antibody vectors for LY-COV555 were used (*31*). For antibodies where vectors were unavailable (e.g., S309, CB6, REGN10933, REGN10987, COV2-2196, COV2-2130, CT-P59, C144, C135, S2E12) (*12*, *13*, *26*, *27*, *30*, *32*–*36*), published amino acids sequences were used for synthesis and cloning into corresponding pVRC8400 vectors (Genscript) (*56*, *57*). For antibody expression, equal amounts of heavy and light chain plasmid DNA were transfected into using Expi293 cells (Gibco, #A14527 by using Expi293 transfection reagent (Gibco, #A14525). The transfected cells were cultured in shaker incubator at 120 rpm, 37°C, 9% CO_2 for 4 to ~5 days. Culture supernatants were harvested and filtered, mAbs were purified over Protein A (Cytiva, #GE17-1279-03) columns. Each antibody was eluted with IgG elution buffer (Pierce, #21009) and immediately neutralized with one tenth volume of 1M Tris-HCL pH 8.0. The antibodies were then buffer exchanged as least twice in PBS by dialysis.

Full-length S constructs

Codon optimized cDNAs encoding full-length S from SARS CoV-2 (GenBank ID: QHD43416.1) were synthesized, cloned into the mammalian expression vector VRC8400 (*56*, *57*) and confirmed by sequencing. S containing D614G amino acid change was generated using the wt S sequence. Other variants containing single or multiple aa changes in the S gene from the S wt or D614G were made by mutagenesis using QuickChange lightning Multi Site-Directed Mutagenesis Kit (Agilent, #210515,) or via synthesis and cloning (Genscript). The S variants tested are B.1.351 (L18F, D80A, D215G, (L242-244)del, R246I, K417N, E484K, N501Y, A701V), P.1 (L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, D614G, H655Y, T1027I, V1176F), B.1.1.7 (H69del, V70del, Y144del, N501Y, A570D, D614G, P681H, T716I, S982A, D1118H), B.1.617.2 (T19R, G142D, E156del, F157del, R158G, L452R, T478K, D614G, P681R, D950N), B.1.1.529 (A67V, H69del, V70del, T95I, G142D, V143del, Y144del, Y145del, N211del, L212I, ins214EPE, G339D, S371L, S373P, S375F, K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, Y505H, T547K, D614G, H655Y, N679K, P681H, N764K, D796Y, N856K, Q954H, N969K, L981F). The S genes containing single RBD amino acid changes from the B.1.1.529 variant were generated based on D614G construct by mutagenesis. These full-length S plasmids were used for pseudovirus production and for cell surface binding assays.

Generation of 293 Flpin-TMPRSS2-ACE2 cell line

293 Flpin-TMPRSS2-ACE2 isogenic cell line was prepared by co-transfecting pCDNA5/FRT plasmid encoding TMPRSS2-T2A-ACE2 and pOG44 plasmid encoding Flp recombinase in 293 Flpin parental cell line (Thermo Fisher, #R75007). Cells expressing TMPRSS2-ACE2 were selected using Hygromycin (Thermo Fisher, #10687010) at 100 micrograms/ml. TMPRSS2 and ACE2 expression profiles in 293 Flpin-TMPSS2-ACE2 were characterized by flow cytometry using a mouse monoclonal antibody against TMPRSS2 (MilliporeSigma, #MABF2158-100UG) followed by an anti-mouse IgG1 APC conjugate (Jackson Laboratories, #115135164) and a molecular probe containing the SARS-CoV-2 receptor binding domain tagged with biotin (Sino Biological, $40592-V08B-B) followed by staining with a BV421 conjugated streptavidin probe (BD Biosciences, #405225).

Pseudovirus neutralization assay

S-containing lentiviral pseudovirions were produced by co-transfection of packaging plasmid pCMVdR8.2, transducing plasmid pHR' CMV-Luc, a TMPRSS2 plasmid and S plasmids from SARS CoV-2 variants into 293T cells using Lipofectamine 3000 transfection reagent (ThermoFisher Scientific, # L3000-001) (*58*, *59*). 293T-ACE2 cells (provided by Dr. Michael Farzan) or 293 flpin-TMPRSS2-ACE2 cells were plated into 96-well white/black Isoplates (PerkinElmer, #6005068) at 75,00 cells per well the day before infection of SARS CoV-2 pseudovirus. Serial dilutions of mAbs were mixed with titrated pseudovirus, incubated for 45 minutes at 37°C and added to cells in triplicate. 293 flpin-TMPRSS2-ACE2 cells were used for some of Class III antibodies like S309 and COV2-2130 while 293T-ACE2 cells were used for the rest of antibodies. Following 2 hours of incubation, wells were replenished with 150 ml of fresh media. Cells were lysed 72 hours later, and luciferase activity was measured with Microbeta (Perking Elmer, #2450-0120). Percent neutralization and neutralization IC_{50} and IC_{80} were calculated using GraphPad Prism 8.0.2.

Cell surface binding

HEK293T cells were transiently transfected with plasmids encoding full length SARS CoV-2 spike variants using lipofectamine 3000 (ThermoFisher, # L3000-001) following manufacturer's protocol. After 40 hours, the cells were harvested and incubated with monoclonal antibodies (0.5 μg/ml) or biotinylated-human ACE2 (Acro Biosystems, AC2-H82F9) for 30 minutes. After incubation with the antibodies or ACE2, the cells were washed and incubated with an allophycocyanin conjugated anti-human IgG (Jackson Immunoresearch Laboratories, #709-136-149) or BV421 conjugated streptavidin conjugate for another 30 minutes. The cells were then washed and fixed with 1% paraformaldehyde (Electron Microscopy Sciences, #15712-S). The samples were then acquired in a BD LSRFortessa X-50 flow cytometer (BD biosciences) and analyzed using Flowjo (BD biosciences). The concentration of ACE2, 10 μg/ml, was determined empirically by titration on WA-1 spike expressing cells. Spike expression level was determined using the SARS-CoV-2 S2 antibody, WS6, that binds to a conserved epitope in the stem-helix in each of the variants (*60*). A variant spike protein expression adjustment factor ($A_{variant}$) was calculated by dividing the mean fluorescent intensity (MFI) for WS6 antibody binding of a variant S by the MFI of WS6 binding to D614G S. Relative binding for antibodies or ACE2 was calculated with the following formula

$$\text{Ligand relative binding} = \frac{A_{\text{variant}} \times (\text{MFI ligand to variant})}{A_{\text{D614G}} \times (\text{MFI ligand to variant})} \times 100\%$$

where ligand is an antibody or ACE2 and variant is D614G, B.1.1.7, B.1.351, P.1, B.1.617.2 or B.1.1.529. The adjustment factor for the sham transfected cells was set to 1.

Production of Fab fragments from monoclonal antibodies

To generate mAb-Fab, IgG was incubated with HRV3C protease (EMD Millipore, #71493) at a ratio of 100 units per 10 mg IgG with HRV 3C Protease Cleavage Buffer (150 mM NaCl, 50 mM Tris-HCl, pH 7.5) at 4°C overnight. Fab was purified by collecting flow-through from Protein A column (GE Health Science), and Fab purity was confirmed by SDS-PAGE.

Determination of binding kinetics of ACE2

Binding kinetics and affinities of ACE2 to SARS-CoV-2 S2P variants were assessed by surface plasma resonance on a Biacore S-200 (GE Healthcare) at 25°C in the HBS-EP+ buffer (10 mM HEPES, pH 7.4, 150 mM NaCl, 3 mM EDTA, and 0.05% surfactant P20). Fc-reactive anti-human IgG antibody (Cytiva, #BR100839) was coupled to a CM5 chip to approximately 10,000 RU, and dimeric, Fc-tagged ACE2 (ACRO Biosystems, AC2-H82F9) at 35 μg/ml was captured for 60 seconds at 10 μl/min to a response of approximately 200 RU. Serially diluted SARS-CoV-2 S2P variants starting at 100 nM were flowed through the sample and reference channels for 180 seconds at 30 μl/min, followed by a 300 second dissociation phase at 30 uL/min. The chip was regenerated using 3 M $MgCl_2$ for 30 seconds at 50 μl/min. Blank sensorgrams were obtained with HBS-EP+ buffer. Blank-corrected sensorgrams of the S2P concentration series were fitted globally with Biacore S200 evaluation software using a 1:1 model of binding. Plots were generated using GraphPad Prism.

Cryo-EM specimen preparation and data collection

Cryo-EM grids for the B.1.1.529 spike stabilized with the "2P" mutations were prepared at 0.5 mg/ml in a buffer containing 10 mM HEPES, pH 7.5 and 150 mM NaCl. For the spike-Fab complexes, the stabilized SARS-CoV-2 spikes of B.1.1.529 or WA-1were1. were mixed with Fab or Fab combinations at a molar ratio of 1.2 Fab per protomer in PBS with final spike protein concentration at 0.5 mg/ml. n-Dodecyl β-D-maltoside (DDM) detergent was added to the protein complex mixtures shortly before vitrification to a concentration of 0.005%. Quantifoil R 2/2 gold grids were subjected to glow discharging in a PELCO easiGlow device (air pressure: 0.39 mBar, current: 20 mA, duration: 30 s) immediately before specimen preparation. Cryo-EM grids were prepared using an FEI Vitrobot Mark IV plunger with the following settings: chamber temperature of 4°C, chamber humidity of 95%, blotting force of –5, blotting time of 2 to 3.5 s, and drop volume of 2.7 µl. Datasets were collected at the National CryoEM Facility (NCEF), National Cancer Institute, on a Thermo Scientific Titan Krios G3 electron microscope equipped with a Gatan Quantum GIF energy filter (slit width: 20 eV) and a Gatan K3 direct electron detector (table S2). Four movies per hole were recorded in the counting mode using Latitude software. The dose rate was 14.65 e-/s/pixel.

Cryo-EM data processing and model fitting

Data process workflow, including motion correction, CTF estimation, particle picking and extraction, 2D classification, ab initio reconstruction, homogeneous refinement, heterogeneous refinement, non-uniform refinement, local refinement and local resolution estimation, were carried out with C1 symmetry in cryoSPARC 3.3 (*61*). The overall resolution was 3.29 Å for the map of B.1.1.529 spike alone structure, 3.85 Å for the map of B.1.1.529 spike in complex with A19-46.1, 2.83 Å for the map of WA-1 spike in complex with A19-61.1 and B1-182.1, and 3.86Å for the map of B.1.1.529 spike in complex with A19-46.1 and B1-182.1. The coordinates of the SARS-CoV-2 spike and Fab B1-182.1 in PDB ID: 7MM0 were used as initial models for fitting the cryo-EM maps. Outputs from AlphaFold 2.0 modelling were used as initial models for Fab A19-46.1 and Fab A19-61.1. To resolve the RBD-antibody interface, local refinements were performed, a mask for the entire spike-antibody complex without the RBD-antibody region was used to extract the particles and a mask encompassing the RBD-antibody region was used for refinement. Local refinements of the Fab A19-46.1 and B.1.1.529 RBD interface and the Fab A19-46.1, Fab B1-182.1 and B.1.1.529 RBD interface resulted 4.68 Å and 4.83 Å maps, respectively, which enabled the definition of the backbone. However, the side chains were not fully resolved. Iterative manual model building and real-space refinement were carried out in Coot (*48*) and in Phenix (*62*), respectively. Molprobity (*63*) was used to validate geometry and check structure quality at each iteration step. UCSF Chimera and ChimeraX were used for map fitting and manipulation (*64*).

Differential scanning calorimetry (DSC)

DSC measurements were performed using a VP-ITC (Microcal) instrument. Spike samples were diluted to 0.125mg/ml in PBS and scanned from 20 to 95°C at a rate of 1°C per minute. Thermal denaturation (T_m) temperature and total enthalpy of unfolding was calculated using the Microcal analysis system in Origin.

Biolayer interferometry binding assay

The antibody binding panel was performed on a FortéBio Octet HTX instrument with black, tilted 384-well plates (Greiner Bio-One). All steps of pre-soaking, binding and dissociation were performed in PBS with 1% BSA at pH 7.4. IgGs and dACE2-Fc were loaded onto Anti-Human Fc Sensor Tips (FortéBio) at a concentration of 1-4µg/ml, resulting in a load response of 0.85-1.5 nm. The plates were agitated at 1,000 rpm and the experiment run at 30°C. Antibodies and ACE2 were loaded onto the tips for 2 minutes, bound to 100nM S2P protein for 5 minutes and dissociated in buffer for 5 minutes. Reference well subtraction was performed with the Data Analysis Software HT v12.0 (FortéBio). The graphs were generated in GraphPad Prism.

REFERENCES AND NOTES

1. F. Wu *et al*., A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020). doi: 10.1038/s41586-020-2008-3; pmid: 32015508
2. Center for Systems Science and Engineering at Johns Hopkins University, Johns Hopkins University COVID-19 Dashboard (2021); https://coronavirus.jhu.edu/map.html.
3. S. Cele *et al*., NGS-SA, COMMIT-KZN Team, A. von Gottberg, J. Bhiman, R. J. Lessells, M.-Y. S. Moosa, M. Davenport, T. de Oliveira, P. L. Moore, A. Sigal, Omicron extensively but incompletely escapes Pfizer BNT162b2 neutralization. *Nature* **602**, 654–656 (2022).
4. World Health Organization (WHO), *Enhancing Readiness for Omicron (B.1.1.529).Technical Brief and Priority Actions for Member States* (WHO, 2021).
5. D. Pinto *et al*., Broad betacoronavirus neutralization by a stem helix-specific human antibody. *Science* **373**, 1109–1116 (2021). doi: 10.1126/science.abj3321; pmid: 34344823
6. D. Wrapp *et al*., Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260–1263 (2020). doi: 10.1126/science.abb2507; pmid: 32075877
7. L. Yurkovetskiy *et al*., Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. *Cell* **183**, 739–751.e8 (2020). doi: 10.1016/j.cell.2020.09.032; pmid: 32991842
8. G. Cerutti *et al*., Potent SARS-CoV-2 neutralizing antibodies directed against spike N-terminal domain target a single supersite. *Cell Host Microbe* **29**, 819–833.e7 (2021). doi: 10.1016/j.chom.2021.03.005; pmid: 33789084
9. A. Rosa *et al*., SARS-CoV-2 can recruit a heme metabolite to evade antibody immunity. *Sci. Adv.* **7**, eabg7607 (2021). doi: 10.1126/sciadv.abg7607; pmid: 33888467
10. G. Cerutti *et al*., Neutralizing antibody 5-7 defines a distinct site of vulnerability in SARS-CoV-2 spike N-terminal domain. *Cell Rep.* **37**, 109928 (2021). doi: 10.1016/j.celrep.2021.109928; pmid: 34706271
11. J. Lan *et al*., Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215–220 (2020). doi: 10.1038/s41586-020-2180-5; pmid: 32225176
12. C. O. Barnes *et al*., SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature* **588**, 682–687 (2020). doi: 10.1038/s41586-020-2852-1; pmid: 33045718
13. D. F. Robbiani *et al*., Convergent antibody responses to SARS-CoV-2 in convalescent individuals. *Nature* **584**, 437–442 (2020). doi: 10.1038/s41586-020-2456-9; pmid: 32555388
14. L. Wang *et al*., Ultrapotent antibodies against diverse and highly transmissible SARS-CoV-2 variants. *Science* **373**, eabh1766 (2021). doi: 10.1126/science.abh1766; pmid: 34210892
15. A. Rambaut *et al*., COVID-19 Genomics Consortium UK, Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. *virological* (2020); https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563.
16. H. Tegally *et al*., Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* **592**, 438–443 (2021). doi: 10.1038/s41586-021-03402-9; pmid: 33690265
17. N. R. Faria *et al*., CADDE Genomic Network, Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings *virological* (2021); https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586.
18. F. Naveca *et al*., Phylogenetic relationship of SARS-CoV-2 sequences from Amazonas with emerging Brazilian variants harboring mutations E484K and N501Y in the Spike protein. *virological* (2021); https://virological.org/t/phylogenetic-relationship-of-sars-cov-2-sequences-from-amazonas-with-emerging-brazilian-variants-harboring-mutations-e484k-and-n501y-in-the-spike-protein/585.
19. World Health Organization, "COVID-19 Weekly Epidemiological Update" (2021), pp. 1–23.
20. T. N. N. Starr *et al*., Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* **182**, 1295–1310.e20 (2020). doi: 10.1016/j.cell.2020.08.012; pmid: 32841599
21. D. Mannar *et al*., SARS-CoV-2 Omicron variant: Antibody evasion and cryo-EM structure of spike protein-ACE2 complex. *Science* **375**, 760–764 (2022). doi: 10.1126/science.abn7760; pmid: 35050643
22. P. Han *et al*., Receptor binding and complex structures of human ACE2 to spike RBD from omicron and delta SARS-CoV-2. *Cell* **185**, 630–640.e10 (2022). doi: 10.1016/j.cell.2022.01.001; pmid: 35093192
23. E. Cameroni *et al*., Broadly neutralizing antibodies overcome SARS-CoV-2 Omicron antigenic shift. *Nature* **602**, 664–670 (2022). doi: 10.1038/s41586-021-04386-2; pmid: 35016195
24. L. Liu *et al*., Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. *Nature* **584**, 450–456 (2020). doi: 10.1038/s41586-020-2571-7; pmid: 32698192
25. W. Yin *et al*., Structures of the Omicron spike trimer with ACE2 and an anti-Omicron antibody. *Science* **375**, 1048–1053 (2022). doi: 10.1126/science.abn8863; pmid: 35133176
26. D. K. Ryu *et al*., The in vitro and in vivo efficacy of CT-P59 against Gamma, Delta and its associated variants of SARS-CoV-2. *Biochem. Biophys. Res. Commun.* **578**, 91–96 (2021). doi: 10.1016/j.bbrc.2021.09.023; pmid: 34547629
27. C. Kim *et al*., A therapeutic neutralizing antibody targeting receptor binding domain of SARS-CoV-2 spike protein. *Nat. Commun.* **12**, 288 (2021). doi: 10.1038/s41467-020-20602-5; pmid: 33436577
28. C. G. Rappazzo *et al*., Broad and potent activity against SARS-like viruses by an engineered human monoclonal antibody. *Science* **371**, 823–829 (2021). doi: 10.1126/science.abf4830; pmid: 33495307
29. K. Westendorf *et al*., LY-CoV1404 (bebtelovimab) potently neutralizes SARS-CoV-2 variants. *bioRxiv* 442182 [Preprint] (2022). doi: 10.1101/2021.04.30.442182
30. M. A. Tortorici *et al*., Ultrapotent human antibodies protect against SARS-CoV-2 challenge via multiple mechanisms. *Science* **370**, 950–957 (2020). doi: 10.1126/science.abe3354; pmid: 32972994

31. B. E. Jones *et al*., The neutralizing antibody, LY-CoV555, protects against SARS-CoV-2 infection in nonhuman primates. *Sci. Transl. Med.* **13**, eabf1906 (2021). doi: 10.1126/scitranslmed.abf1906; pmid: 33820835
32. R. Shi *et al*., A human neutralizing antibody targets the receptor-binding site of SARS-CoV-2. *Nature* **584**, 120–124 (2020). doi: 10.1038/s41586-020-2381-y; pmid: 32454512
33. J. Hansen *et al*., Studies in humanized mice and convalescent humans yield a SARS-CoV-2 antibody cocktail. *Science* **369**, 1010–1014 (2020). doi: 10.1126/science.abd0827; pmid: 32540901
34. S. J. Zost *et al*., Potently neutralizing and protective human antibodies against SARS-CoV-2. *Nature* **584**, 443–449 (2020). doi: 10.1038/s41586-020-2548-6; pmid: 32668443
35. D. Pinto *et al*., Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. *Nature* **583**, 290–295 (2020). doi: 10.1038/s41586-020-2349-y; pmid: 32422645
36. L. Piccoli *et al*., Mapping neutralizing and immunodominant sites on the SARS-CoV-2 spike receptor-binding domain by structure-guided high-resolution serology. *Cell* **183**, 1024–1042.e21 (2020). doi: 10.1016/j.cell.2020.09.037; pmid: 32991844
37. W. Dejnirattisai *et al*., SARS-CoV-2 Omicron-B.1.1.529 leads to widespread escape from neutralizing antibody responses. *Cell* **185**, 467–484.e15 (2022). doi: 10.1016/j.cell.2021.12.046; pmid: 35081335
38. A. J. Greaney *et al*., Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies. *Nat. Commun.* **12**, 4196 (2021). doi: 10.1038/s41467-021-24435-8; pmid: 34234131
39. C. K. Wibmer *et al*., SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. *Nat. Med.* **27**, 622–625 (2021). doi: 10.1038/s41591-021-01285-x; pmid: 33654292
40. P. Wang *et al*., Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. *Nature* **593**, 130–135 (2021). doi: 10.1038/s41586-021-03398-2; pmid: 33684923
41. A. Baum *et al*., Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. *Science* **369**, 1014–1018 (2020). doi: 10.1126/science.abd0831; pmid: 32540904
42. P. Wang *et al*., Increased resistance of SARS-CoV-2 variant P.1 to antibody neutralization. *Cell Host Microbe* **29**, 747–751.e4 (2021). doi: 10.1016/j.chom.2021.04.007; pmid: 33887205
43. T. N. Starr, A. J. Greaney, A. S. Dingens, J. D. Bloom, Complete map of SARS-CoV-2 RBD mutations that escape the monoclonal antibody LY-CoV555 and its cocktail with LY-CoV016. *Cell Rep. Med.* **2**, 100255 (2021). doi: 10.1016/j.xcrm.2021.100255; pmid: 33842902
44. K. G. Nabel *et al*., Structural basis for continued antibody evasion by the SARS-CoV-2 receptor binding domain. *Science* **375**, eabl6251 (2022). doi: 10.1126/science.abl6251; pmid: 34855508
45. J. Dong *et al*., Genetic and structural basis for recognition of SARS-CoV-2 spike protein by a two-antibody cocktail. *bioRxiv* 428529 [Preprint] (2021). doi: 10.1101/2021.01.27.428529
46. S. H. Ko *et al*., High-throughput, single-copy sequencing reveals SARS-CoV-2 spike variants coincident with mounting humoral immunity during acute COVID-19. *PLOS Pathog.* **17**, e1009431 (2021). doi: 10.1371/journal.ppat.1009431; pmid: 33831133
47. K. S. Corbett *et al*., Protection against SARS-CoV-2 Beta variant in mRNA-1273 vaccine-boosted nonhuman primates. *Science* **374**, 1343–1353 (2021). doi: 10.1126/science.abl8912; pmid: 34672695
48. L. Liu *et al*., Striking antibody evasion manifested by the Omicron variant of SARS-CoV-2. *Nature* **602**, 676–681 (2022). doi: 10.1038/s41586-021-04388-0; pmid: 35016198
49. S. Iketani *et al*., Antibody evasion properties of SARS-CoV-2 Omicron sublineages. *bioRxiv* 479306 [Preprint] (2022). doi: 10.1101/2022.02.07.479306
50. Y. Cao *et al*., Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. *Nature* **602**, 657–663 (2022). doi: 10.1038/s41586-021-04385-3; pmid: 35016194
51. M. McCallum *et al*., Structural basis of SARS-CoV-2 Omicron immune evasion and receptor engagement. *Science* **375**, 864–868 (2022). doi: 10.1126/science.abn8652; pmid: 35076256
52. L. A. VanBlargan *et al*., An infectious SARS-CoV-2 B.1.1.529 Omicron virus escapes neutralization by therapeutic monoclonal antibodies. *Nat. Med.* **28**, 490–495 (2022). doi: 10.1038/s41591-021-01678-y; pmid: 35046573
53. W. R. Gallaher, Omicron is a multiply recombinant set of variants that have evolved over many months. *virological* (2021); https://virological.org/t/omicron-is-a-multiply-recombinant-set-of-variants-that-have-evolved-over-many-months/775.
54. K. A. Howell *et al*., Cooperativity enables non-neutralizing antibodies to neutralize Ebolavirus. *Cell Rep.* **19**, 413–424 (2017). doi: 10.1016/j.celrep.2017.03.049; pmid: 28402862
55. T. Zhou *et al*., Structure-based design with Tag-based purification and in-process biotinylation enable streamlined development of SARS-CoV-2 spike molecular probes. *Cell Rep.* **33**, 108322 (2020). doi: 10.1016/j.celrep.2020.108322; pmid: 33091382
56. D. H. Barouch *et al*., A human T-cell leukemia virus type 1 regulatory element enhances the immunogenicity of human immunodeficiency virus type 1 DNA vaccines in mice and nonhuman primates. *J. Virol.* **79**, 8828–8834 (2005). doi: 10.1128/JVI.79.14.8828-8834.2005; pmid: 15994776
57. A. T. Catanzaro *et al*., Phase I clinical evaluation of a six-plasmid multiclade HIV-1 DNA candidate vaccine. *Vaccine* **25**, 4085–4092 (2007). doi: 10.1016/j.vaccine.2007.02.050; pmid: 17391815
58. L. Naldini, U. Blömer, F. H. Gage, D. Trono, I. M. Verma, Efficient transfer, integration, and sustained long-term expression of the transgene in adult rat brains injected with a lentiviral vector. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 11382–11388 (1996). doi: 10.1073/pnas.93.21.11382; pmid: 8876144
59. Z. Y. Yang *et al*., Evasion of antibody neutralization in emerging severe acute respiratory syndrome coronaviruses. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 797–801 (2005). doi: 10.1073/pnas.0409065102; pmid: 15642942
60. W. Shi *et al*., Vaccine-elicited murine antibody WS6 neutralizes diverse beta-coronaviruses by recognizing a helical stem supersite of vulnerability. *bioRxiv*2022.01.25.477770 (2022). https://doi.org/.doi: 10.1101/2022.01.25.477770
61. A. Punjani, J. L. Rubinstein, D. J. Fleet, M. A. Brubaker, cryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017). doi: 10.1038/nmeth.4169; pmid: 28165473
62. P. V. Afonine *et al*., Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 352–367 (2012). doi: 10.1107/S0907444912001308; pmid: 22505256
63. I. W. Davis, L. W. Murray, J. S. Richardson, D. C. Richardson, MOLPROBITY: Structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Res.* **32**, W615–W619 (2004). doi: 10.1093/nar/gkh398; pmid: 15215462
64. E. F. Pettersen *et al*., UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004). doi: 10.1002/jcc.20084; pmid: 15264254

ACKNOWLEDGMENTS

We thank N. A. Doria-Rose, W.-P. Kong, S. O'Dell, and S. D. Schmitt for assistance in B.1.1.529 plasmid production and distribution; M. Kanekiyo for cell line assistance; J. Stuckey and S. Wang for assistance with manuscript preparation and submission; and members of the Virology Laboratory, Vaccine Research Center, for discussions and comments on the manuscript. We thank S. Žentelis, E. Lameignere, and K. Westerndorf for antibody LY-CoV1404. We are grateful to T. Edwards and T. L. Fox of NCEF for cryo-EM data collection and for technical assistance with cryo-EM data processing. **Funding:** This work was funded by the Intramural Research Program of the Vaccine Research Center, NIAID, NIH. This research was, in part, supported by the National Cancer Institute's National Cryo-EM Facility at the Frederick National Laboratory for Cancer Research under contract HSSN261200800001E. **Author contributions:** T.Z., L.W., J.M., and A.P. designed experiments and analyzed data. L.W., A.P., Y.Z., D.R.H., C.A.T., A.S.O., E.S.Y., M.Che., K.L., and E.-S.D.S. performed experiments. L.W., J.M., A.S.O., W.S., M.Cho., I-T.T., A.C., T.L., and B.Z. produced proteins, antibodies, and other reagents. T.Z. led electron microscopy studies, assisted by C.J., T.S., and Y.T.; J.M., B.S.G., J.R.M., N.J.S., and P.D.K. supervised experiments. T.Z., L.W., J.M., N.J.S., and P.D.K. wrote the manuscript, with help from all authors. **Competing interests:** T.Z., L.W., J.M., A.P., Y.Z., E.S.Y., W.S., J.R.M., N.J.S., and P.D.K. are inventors on US patent application no. 63/147,419. J.R.M., B.S.G., L.W., Y.Z., and W.S. are inventors on PCT/US2020/063991 and PCT/US2021/020843. **Data and materials availability:** All data are available in the main text or the supplementary materials. Atomic coordinates and cryo-EM maps of the reported structure have been deposited into the Protein Data Bank and Electron Microscopy Data Bank under the session codes PDB 7TB4 and EMD-25792 for the SARS-CoV-2 B.1.1.529 VOC spike, PDB 7TCA and EMD-25807 for SARS-CoV-2 B.1.1.529 VOC spike in complex with antibody A19-46.1, PDB 7TC9 and EMD-25806 for local refinement of the SARS-CoV-2 B.1.1.529 VOC RBD in complex with antibody A19-46.1, PDB 7TCC and EMD-25808 for SARS-CoV-2 B.1.1.529 VOC spike in complex with antibodies B1-182.1 and A19-46.1, PDB 7U0D and EMD-26256 for local refinement of the SARS-CoV-2 B.1.1.529 VOC RBD in complex with antibodies B1-182.1 and A19-46.1, PDB 7TB8 and EMD-25794 for SARS-CoV-2 WA-1 spike in complex with antibodies A19-61.1 and B1-182.1, and PDB 7TBF and EMD-25797 for local refinement of the SARS-CoV-2 WA-1 RBD in complex with antibodies A19-61.1 and antibody B1-182.1. Original materials in this manuscript are available under a materials transfer agreement with the National Institutes of Health. This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/. This license does not apply to figures/photos/artwork or other content included in the article that is credited to a third party; obtain authorization from the rights holder before using such material.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abn8897
Figs. S1 to S10
Tables S1 and S2
MDAR Reproducibility Checklist

28 December 2021; accepted 19 March 2022
Published online 24 March 2022
10.1126/science.abn8897

RESEARCH ARTICLE SUMMARY

TRAFFIC SAFETY

Can behavioral interventions be too salient? Evidence from traffic safety messages

Jonathan D. Hall and Joshua M. Madsen*

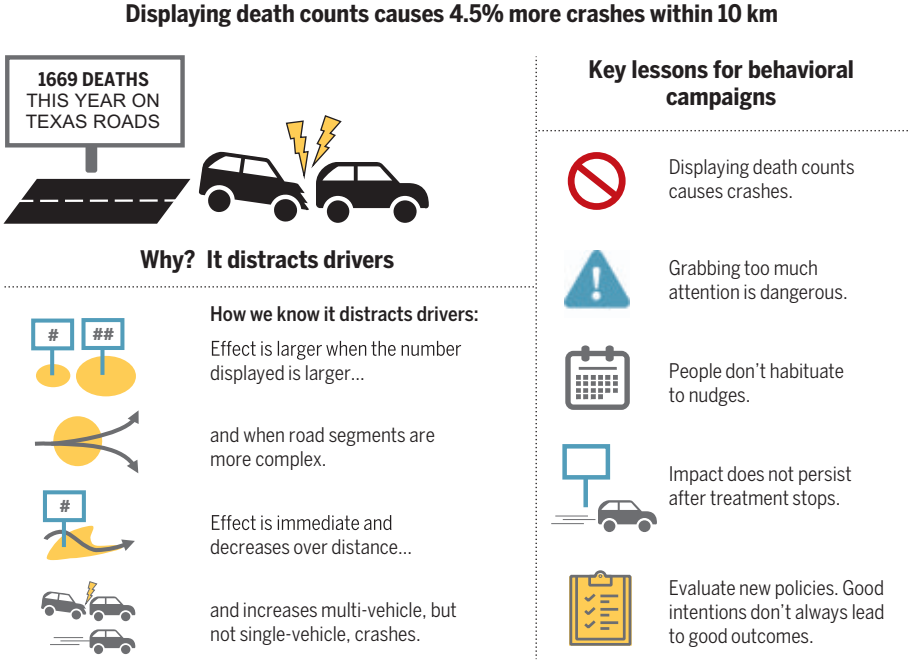
INTRODUCTION: Policy-makers are increasingly turning to behavioral interventions such as nudges and informational campaigns to address a variety of issues. Guidebooks say that these interventions should "seize people's attention" at a time when they can take the desired action, but little consideration has been given to the costs of seizing one's attention and to the possibility that these interventions may crowd out other, more important, considerations. We estimated these costs in the context of a widespread, seemingly innocuous behavioral campaign with the stated objective of reducing traffic crashes. This campaign displays the year-to-date number of statewide roadside fatalities (fatality messages) on previously installed highway dynamic message signs (DMSs) and has been implemented in 28 US states.

RATIONALE: We estimated the impact of displaying fatality messages using data from Texas. Texas provides an ideal setting because the Texas Department of Transportation (TxDOT) decided to show fatality messages starting in August 2012 for 1 week each month: the week before TxDOT's monthly board meeting (campaign weeks). This allows us to measure the impact of the intervention, holding fixed the road segment, year, month, day of week, and time of day. We used data on 880 DMSs and all crashes occurring in Texas between 1 January 2010 and 31 December 2017 to investigate the effects of this safety campaign. We estimated how the intervention affects crashes near DMSs as well as statewide. As placebo tests, we estimated whether the chosen weeks inherently differ using data from before TxDOT started displaying fatality messages and data from upstream of DMSs.

RESULTS: Contrary to policy-makers' expectations, we found that displaying fatality messages increases the number of traffic crashes. Campaign weeks realize a 1.52% increase in crashes within 5 km of DMSs, slightly diminishing to a 1.35% increase over the 10 km after DMSs. We used instrumental variables to recover the effect of displaying a fatality message and document a significant 4.5% increase in the number of crashes over 10 km. The effect of displaying fatality messages is comparable to raising the speed limit by 3 to 5 miles per hour or reducing the number of highway troopers by 6 to 14%. We also found that the total number of statewide on-highway crashes is higher during campaign weeks. The social costs of these fatality messages are large: Back-of-the-envelope calculations suggest that this campaign causes an additional 2600 crashes and 16 fatalities per year in Texas alone, with a social cost of $377 million per year.

Our proposed explanation for this surprising finding is that these "in-your-face," "sobering," negatively framed messages seize too much attention (i.e., are too salient), interfering with drivers' ability to respond to changes in traffic conditions. Supporting this explanation, we found that displaying a higher fatality count (i.e., a plausibly more attention-grabbing statistic) causes more crashes than displaying a small one, that fatality messages are more harmful when displayed on more complex road segments, that fatality messages increase multi-vehicle crashes (but not single-vehicle crashes), and that the impact is largest close to DMSs and decreases over longer distances. We discuss seven alternative hypotheses, including the possibilities that treated weeks are inherently more dangerous and that fatality messages help in the long run. We provide evidence inconsistent with each alternative hypothesis.

CONCLUSION: Our study highlights five key lessons. First, and most directly, fatality message campaigns increase the number of crashes, so ceasing these campaigns is a low-cost way to improve traffic safety. Second, behavioral interventions can be too salient, crowding out more essential considerations and causing the intervention to backfire with costly consequences. Thus the message, delivery, and timing of behavioral interventions should be carefully designed so they are not too salient relative to individuals' cognitive loads when the intervention occurs. Third, individuals don't necessarily habituate to behavioral interventions, even after years of treatment. Fourth, the effects of interventions do not necessarily persist after treatment stops. Finally, it is important to measure an intervention's effect, even for simple interventions, because good intentions do not necessarily imply good outcomes. ■



A traffic safety campaign that leads to more crashes.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: jmmadsen@umn.edu
Cite this article as J. D. Hall *et al.*, *Science* **376**, eabm3427 (2022). DOI: 10.1126/science.abm3427

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abm3427

RESEARCH ARTICLE

TRAFFIC SAFETY

Can behavioral interventions be too salient? Evidence from traffic safety messages

Jonathan D. Hall[1,2] and Joshua M. Madsen[3]*

Although behavioral interventions are designed to seize attention, little consideration has been given to the costs of doing so. We estimated these costs in the context of a safety campaign that, to encourage safe driving, displays traffic fatality counts on highway dynamic message signs for 1 week each month. We found that crashes increase statewide during campaign weeks, which is inconsistent with any benefits. Furthermore, these effects do not persist beyond campaign weeks. Our results show that behavioral interventions, particularly negatively framed ones, can be too salient, crowding out more important considerations and causing interventions to backfire—with costly consequences.

There is growing interest among academics and policy-makers in using behavioral interventions as a low-cost and easy-to-implement way of encouraging socially desirable behaviors. Reflecting this interest, such interventions are now used by >200 governments and institutions worldwide to address a variety of issues, including voter turnout, charitable giving, retirement savings, water conservation, energy conservation, hand washing, caloric intake, diarrhea, and risky sexual behavior (*1*–*3*). Many of these interventions are expressly designed to "seize people's attention" at a time when they can make the desired action (*4*), a characteristic that we refer to as salience (*5*–*7*). However, little consideration has been given to individuals' cognitive constraints and to the possibility that seizing one's attention may crowd out other, more important considerations (such as focusing on the task at hand).

Our context is a seemingly innocuous behavioral campaign with the stated objective of reducing traffic crashes, the leading cause of death for 5- to 45-year-olds in the United States and worldwide (*8*, *9*). This campaign displays the year-to-date count of statewide roadside fatalities on previously installed dynamic message signs (DMSs) (e.g., "1669 deaths this year on Texas roads"; fig. S1). These fatality messages are expressly designed to be salient, with official statements expressing the "hope" that these "in-your-face" safety messages will "motivate motorists to exercise caution behind the wheel" and that a "sobering new message ... will [hopefully] help save lives" (*10*, *11*). Because of its low cost and ease of implementation, this campaign has spread to at least 28 US states since 2012 and affected >100 million drivers (*12*).

This campaign is widely believed to be effective. For instance, in Illinois, the decision to start displaying fatality messages was unanimously supported by the Department of Transportation, the State Police, and the Department of Public Health (*13*). Many drivers also believe that fatality statistics make safety messages more effective (*14*, *15*). Belief in the effectiveness of these messages is likely a factor in their rapid spread.

One key challenge when measuring the effect of fatality messages on crashes is that they are frequently displayed during safer times, when the DMS is not being used for more pressing concerns (e.g., travel times and crash alerts), biasing any naïve analysis toward finding a lower frequency of crashes when fatality messages are displayed (*16*).

The State of Texas provides a unique setting in which to overcome this challenge. Unlike most states, the Texas Department of Transportation (TxDOT) displays the current statewide fatality count only 1 week each month: the week before TxDOT's monthly board meeting. Although more important messages regularly preempt the fatality message, traffic engineers are instructed that along corridors with a large number of DMSs, "the fatality message should be displayed on a few [DMSs]" (fig. S2). We confirmed that fatality messages concentrate in the designated weeks and used this assignment to treatment to estimate the effect of fatality messages on traffic crashes.

We estimated the effect of displaying fatality messages relative to the status quo usage of DMSs by comparing how the number of crashes downstream of a DMS differs the week before a TxDOT board meeting ("campaign week") relative to the same road segment the rest of the month. We conducted our analysis at the segment-hour level and included an extensive fixed-effect structure to control for inherent variation in crash risk across different segments over time and throughout each day. As such, our estimates compare, for example, the number of crashes within 10 km downstream of a DMS from 2:00 to 3:00 p.m. on Thursday, 18 July 2013 (which occurred during the week before a board meeting) against the number of crashes on the same road segment from 2:00 to 3:00 p.m. on the other three Thursdays in July 2013.

We conducted two tests to address the possibility that the weeks before TxDOT board meetings are inherently more dangerous than other weeks within the same month. First, we estimated the change in crashes occurring upstream of DMSs. Second, we estimated a placebo effect using data from before TxDOT began displaying fatality messages.

Our main results are difference-in-differences estimates that exploit both within-month variation in when fatality messages are instructed to be displayed (campaign weeks versus other weeks) and differences between the pretreatment (January 2010 to July 2012) and treatment (August 2012 to December 2017) periods.

Results

We begin with univariate analyses documenting an increase in crashes the week before a board meeting (when fatality messages are displayed) relative to other weeks. We then show that these results hold in first-difference and difference-in-differences multivariate analyses after controlling for weather, holidays, and segment-year-month-time-of-day-day-of-week fixed effects. We conclude by estimating the impact of displaying a fatality message using instrumental variables and show that the impact of campaign weeks has not dissipated over time.

Univariate results

Figure 1 shows that there are more crashes downstream of DMSs during campaign weeks than in other weeks. Specifically, the circles plot the percentage difference in the average number of crashes occurring during campaign weeks versus other weeks over the segments [0,1], (1–4], (4–7], and (7–10] km downstream of DMSs. We found that there are more crashes during campaign weeks, with the largest effect being a 2.7% increase over the first kilometer (P = 0.012). This effect diminishes to a 1.8% increase over the (7–10] km interval (P < 0.001).

The results shown in Fig. 1 suggest that although the estimated effect diminishes over longer distances, it does not decay to 0. We conjectured that the increase in crashes over distances farther away from DMSs is caused by subsequent treatment by downstream DMSs. To map out the impact of fatality messages in

[1]Department of Economics and Munk School of Global Affairs and Public Policy, University of Toronto, Toronto, Ontario M5S 3G7, Canada. [2]Department of Spatial Economics, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, Netherlands. [3]Carlson School of Management, University of Minnesota, Minneapolis, MN 55455, USA.
*Corresponding author. Email: jmmadsen@umn.edu

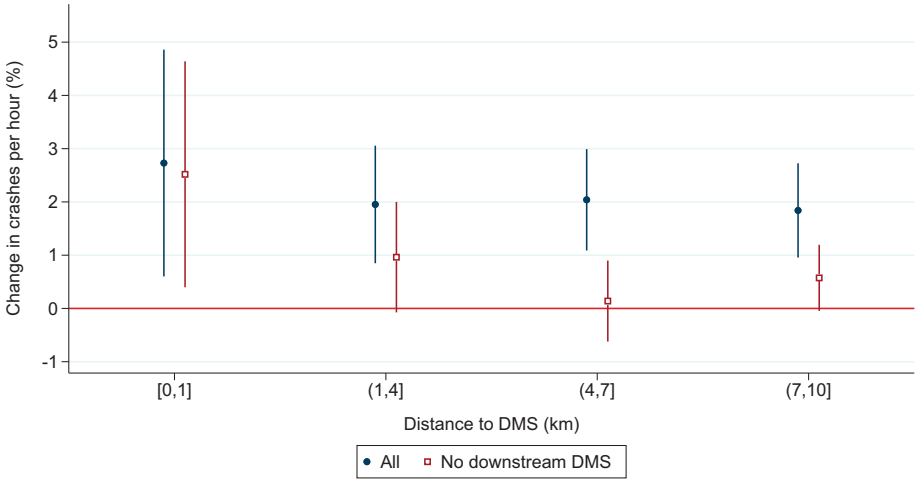


Fig. 1. Effect of fatality message on crashes by distance from DMS: Univariate. Shown is the percent change in the number of crashes on Texas highways during weeks that precede TxDOT board meetings (campaign weeks) relative to all other weeks. Highway crashes are measured over hour *h* of day *d* over distance *x* (relative to DMS *s*) and are indicated on the *x*-axis. The circles plot the difference in the average number of crashes during campaign weeks and all other hours and the associated 95% confidence intervals (bars). The hollow squares plot the difference in the average number of crashes for the sample of DMSs with no downstream DMS within *x* km; that is, for the distance (1,4], the closest downstream DMS is ≥4 km away. We scaled crash counts by the population average for all segments of the same distance *x* and multiplied by 100. Standard errors are clustered by geography-year-month, where geography indicates a bin of size x^2 km^2 containing the DMS. The sample period is August 2012 to December 2017.

the absence of subsequent treatments, the hollow squares in Fig. 1 plot univariate differences in crash rates for the subset of DMSs in which there are no downstream DMSs within *x* km. We found that for DMSs with no downstream DMS within 7 or 10 km, the effect over distances (4,7] and (7,10] km becomes statistically insignificant, respectively. These results suggest that the immediate increase in crashes in response to the fatality message is short lived and concentrated after DMSs.

We next conducted two placebo tests to address the possibility that the week before board meetings is inherently more dangerous than other weeks. First, we examined the change in crashes upstream of DMSs. Because a segment upstream of one DMS may be downstream of another, we limited this test to DMSs in which the nearest upstream DMS is >10 km away (reducing our sample by 75%). As shown in fig. S3A, we found no upstream effect for this restricted sample. All but one of the downstream estimates is >0, with a 1.7% increase over the (7–10] km downstream of DMSs (P = 0.018). The lack of a significant upstream effect for this subsample of DMSs is consistent with fatality messages driving the increase in crashes, although we caution that we have less power to measure an effect on this nonrandom sample of DMSs, which mostly include DMSs on the edge of cities or in rural areas. Second, we estimated the change in crashes during the week before a TxDOT board meeting for the pretreatment period. As shown in fig. S3B, we found neither a downstream effect nor a positive upstream effect during this period.

Multivariate results

We next show that these results hold when using more rigorous specifications that adjust for weather, holidays, and segment-year-month-time-of-day-day-of-week fixed effects. We start with first-difference estimates, finding that the more rigorous specification reduces the treatment effect by up to 50% (fig. S4). Multivariate versions of our two placebo tests produce similar results (fig. S5).

Table 1 reports our main results: difference-in-differences estimates that account for the uncertainty in whether the week before a board meeting is inherently more dangerous. Each column in Table 1 reports results for different highway segment lengths. The first row, "campaign week × post," estimates the treatment effect of fatality messages. We found that within 5 km of a DMS, there is a 1.52% increase in the number of crashes per hour (P = 0.025), slightly diminishing to a 1.35% increase over the 10 km after DMSs (P = 0.025). Within 3 km, the effect is positive but not statistically significant. The second row, "campaign week," estimates the change in the number of crashes 1 week before board meetings from January 2010 to August 2012. Because this period predates the fatality safety campaigns, we expected and indeed found no effect (consistent with fig. S5B); these estimates are both small and statistically insignificant. table S2 presents similar findings separately analyzing both pretreatment and treatment periods.

Our estimated magnitudes are large given the intervention's simplicity and are estimates of the impact of campaign weeks on crashes. Because of imperfect compliance and competing demands, traffic engineers do not display fatality messages on all DMS hours during campaign weeks, implying that the effect of displaying a fatality message on a DMS is even larger. We used two-sample instrumental variables to estimate the effect of displaying a fatality message on the number of crashes downstream of a DMS. The first-stage regression was run on the subsample for which we have DMS log files, and the second-stage regression was run on the full sample. We bootstrapped standard errors. The first-stage results are presented in table S3, and the second-stage results are presented in table S4. We found that displaying a fatality message results in a positive but insignificant increase in crashes over the first 3 km. Consistent with our earlier results, we found a larger and statistically significant 5% increase in the number of crashes over 5 km and a slightly smaller increase of 4.5% over 10 km when fatality messages are displayed. These magnitudes are comparable to increasing the speed limit by 3 to 5 miles per hour (*17*) or reducing the number of highway troopers by 6 to 14% (*18*).

We found no evidence that the effect of displaying a fatality message has dissipated over time. Figure 2 plots coefficient estimates when the treatment effect is allowed to vary each year, using 2011 as the base year. We found that the treatment effect does not change from 2010 to 2012 (generally the pretreatment period). For all years after 2012 except 2016, the estimated coefficient is positive, with P ranging from 0.018 to 0.068 (*19*).

As described in the supplementary text, section S2, we found no evidence that the types of vehicles or drivers (by age and gender) involved in crashes differ during campaign weeks.

Mechanism

We have shown that displaying fatality messages on DMSs increases the number of crashes. In this section, we investigate the mechanism for this increase. A large body of research documents that attention and working memory are scarce resources, and that distractions create extraneous cognitive load that hampers individuals' ability to process new information (*7*, *20*–*22*). Examples include longer response times, more mistakes, and failure to process available information (*23*–*25*). Fatality messages plausibly add greater cognitive load than a typical DMS message because they are designed to be more salient than the typical message and (intentionally) communicate that

Table 1. Effect of fatality messages on crashes. Shown are estimates of the effect of campaign weeks on traffic crashes. The sample period is 1 January 2010 through 31 December 2017. The dependent variable is the number of crashes occurring on highway segment s of length x km on date d during hour h, scaled by the population average for all segments of length x and multiplied by 100. Highway segments begin at each DMS located on a highway and continue for x km of highway driving distance, where $x \in \{3,5,10\}$, and are denoted in the column headers. We used as our primary right-side variable campaign week$_{d,h}$, an indicator variable for whether day d and hour h fell within a campaign week. The variable post$_d$ indicates observations after 1 August 2012. We include, but do not tabulate, indicators for whether either trace precipitation or more than trace precipitation was measured on segment s during hour h, using data from the closest weather station (trace precipitation$_{s,d,h}$ and precipitation$_{s,d,h}$, respectively) and interactions between these measures and post$_d$. We also include segment-year-month-day-of-week-hour (S-Y-M-D-H) and holiday fixed effects (FE). Standard errors are clustered by geography-year-month and are shown in parentheses, where geography indicates a bin of size x^2 km^2 containing the DMS. **$P < 0.05$. The equation used was as follows, where dow(d) is the day of the week associated with day d: $\text{crash}(\%)_{s(x),d,h} = \delta \cdot \text{campaign week}_{d,h} \cdot \text{post}_d + \beta_1 \cdot \text{campaign week}_{d,h} + \beta_2 \cdot \text{trace precipitation}_{s,d,h} + \beta_3 \cdot \text{trace precipitation}_{s,d,h} \cdot \text{post}_d + \beta_4 \cdot \text{precipitation}_{s,d,h} + \beta_5 \cdot \text{precipitation}_{s,d,h} \cdot \text{post}_d + \gamma_{s,m(d),dow(d),h} + \zeta_{\text{holiday}} + \varepsilon_{s,d,h}$

	Crashes per hour (%)		
	3 km	**5 km**	**10 km**
	(1)	(2)	(3)
Campaign week × post	1.13 (0.86)	1.52 (0.68)**	1.35 (0.60)**
Campaign week	0.35 (0.63)	−0.27 (0.48)	−0.32 (0.43)
Observations	61,697,666	61,697,666	61,697,666
Adjusted R^2	0.02	0.03	0.08
Rain and interactions	Yes	Yes	Yes
S-Y-M-D-H FE	Yes	Yes	Yes
Holiday FE	Yes	Yes	Yes

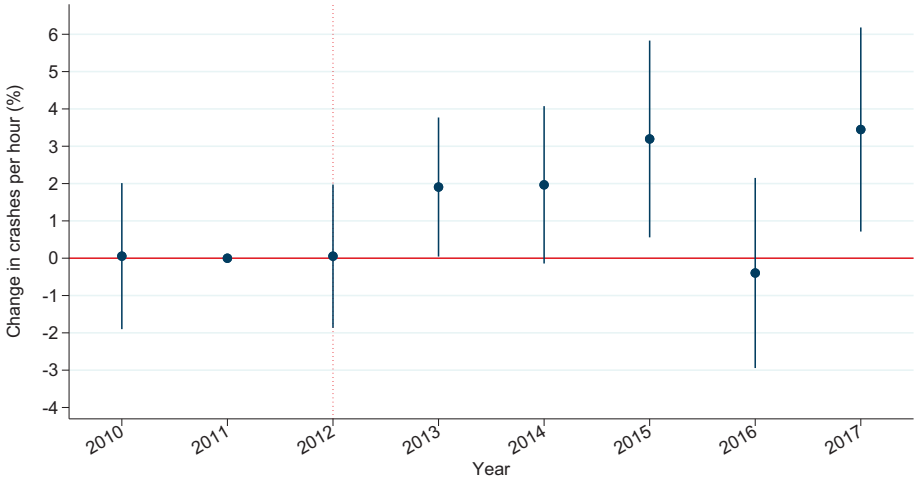


Fig. 2. Effect of fatality messages on crashes by year. Shown are the δ_i coefficient estimates (circles) and the associated 95% confidence intervals (bars) from the regression below that allows the treatment effect to vary by year. Treatment effects are estimated relative to the treatment effect in 2011. The dependent variable, $\text{crash}(\%)_{s(10),d,h}$, is the scaled number of crashes occurring on day d during hour h over the 10 km downstream of DMS s; campaign week$_{d,h}$ is an indicator variable for whether day d and hour h fell within a campaign week; and year$_{d,i}$ is an indicator variable if day d was in year i. Standard errors are clustered by geography-year-month bins, where geography bins are defined as the 10^2 km^2 containing the DMS. The dotted vertical line indicates that treatment started in August 2012. The equation used was as follows, where dow(d) is the day of the week associated with day d: $\text{crash}(\%)_{s(10),d,h} = \sum_{i \in \{2010,\ 2012,\ \ldots,\ 2017\}} \delta_i \cdot \text{campaign week}_{d,h} \cdot \text{year}_{d,i} + \beta_1 \cdot \text{campaign week}_{d,h} + \beta_2 \cdot \text{trace precipitation}_{s,d,h} + \beta_3 \cdot \text{precipitation}_{s,d,h} + \gamma_{s,m(d),dow(d),h} + \zeta_{\text{holiday}} + \varepsilon_{s,d,h}$

driving can be deadly (i.e., a negatively framed message). Relative to other messages, fatality messages may thus add more to drivers' cognitive loads by inducing anxiety about death. Psychologists have documented that high levels of anxiety or arousal can worsen performance on a variety of tasks by causing individuals to focus on the risk rather than on the task and causing some to overthink their actions, overriding faster automatic responses (*26*, *27*). We thus hypothesized that fatality messages temporarily increase drivers' cognitive loads, reducing their ability to safely and quickly respond to changes in traffic conditions (e.g., stay in lane, maintain proper distance, respond to a vehicle changing lanes) and making them more likely to be involved in a crash. For conciseness, we refer to this as "distraction." We provide eight pieces of evidence supporting this hypothesis, focusing on the treatment effect on crashes occurring within 10 km of a DMS.

Our first piece of evidence supporting the distraction hypothesis is that the harm done by the fatality message is larger when the reported number of statewide deaths is larger, suggesting that bigger fatality statistics are more salient and distracting than smaller ones. We estimated a regression that allows the effect of campaign weeks to vary by the quartile of reported deaths. Figure 3 plots these results. When the number of reported deaths is small, displaying a fatality message decreases the number of crashes by 2.8% ($P = 0.024$). However, as the number of reported deaths increases, the effect of displaying a fatality message on crashes grows more harmful, reaching a 5.0% increase ($P = 0.003$).

Second, and closely related, the harm done by the fatality message increases throughout the year. Because the number of deaths reported mechanically climbs throughout the year, showing an increase in crashes throughout the year is an alternate way of showing that increases in the displayed number of deaths lead to more crashes. Figure 4 plots our difference-in-differences estimates of the treatment effect by calendar month. We found that displaying a fatality message in February, when the number of deaths displayed resets (January displays the prior year's total), reduces crashes by 3.4% ($P = 0.113$); this effect then worsens throughout the year. From October through January, the effect is positive and statistically significant. In supplementary text, section S3, we discuss possible reasons why June, July, and August deviate from this pattern.

Third, as Fig. 4 shows, the effect of displaying a fatality message drops 11 percentage points between January and February, when the displayed number of deaths resets. This significant change supports the hypothesis that the number displayed matters and

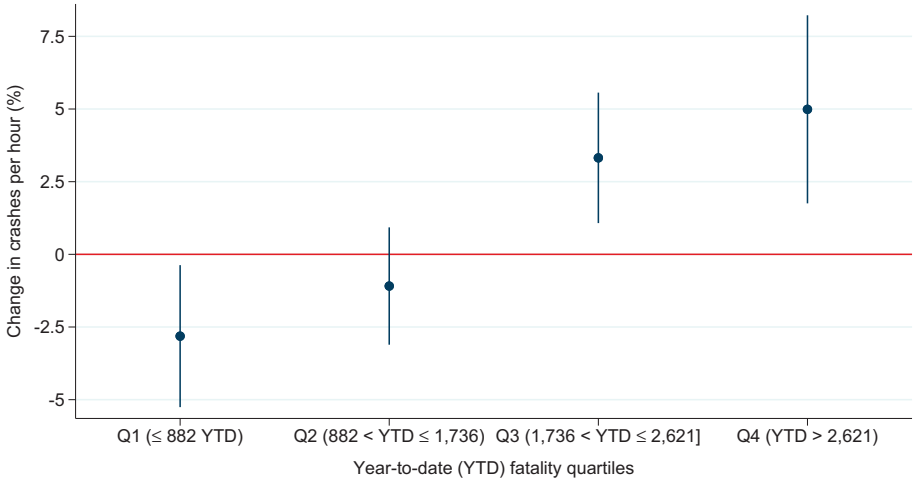


Fig. 3. Effect of fatality messages on crashes by YTD death quartile. Shown are the δ_i coefficient estimates (circles) and the associated 95% confidence intervals (bars) from the regression below that allows the treatment effect to vary by the year-to-date (YTD) number of deaths on Texas roads. $\text{YTDquartile}_{d,i}$ is an indicator if, on day d, the YTD number of deaths was in quartile i, and post_d is an indicator for observations after 1 August 2012. Remaining variables are defined in Fig. 2 (see also table S9). Standard errors are clustered by geography-year-month bins, where geography bins are defined as the 10^2 km^2 containing the DMS. The sample period is January 2010 to December 2017. The equation used was as follows:

$$\text{crash}(\%)_{s(10),d,h} = \sum_{i\in\{\text{quartile1},\ldots,\text{quartile4}\}} \delta_i \bullet \text{campaign week}_{d,h} \bullet \text{YTDquartile}_{d,i} \bullet \text{post}_d + \sum_{i\in\{\text{quartile1},\ldots,\text{quartile4}\}} \beta_{1,i} \bullet \text{campaign week}_{d,h} \bullet \text{YTDquartile}_{d,i} + \beta_2 \bullet \text{trace precipitation}_{s,d,h} + \beta_3 \bullet \text{trace precipitation}_{s,d,h} \bullet \text{post}_d + \beta_4 \bullet \text{precipitation}_{s,d,h} + \beta_5 \bullet \text{precipitation}_{s,d,h} \bullet \text{post}_d + \gamma_{s,m(d),dow(d),h} + \zeta_{\text{holiday}} + \varepsilon_{s,d,h}$$

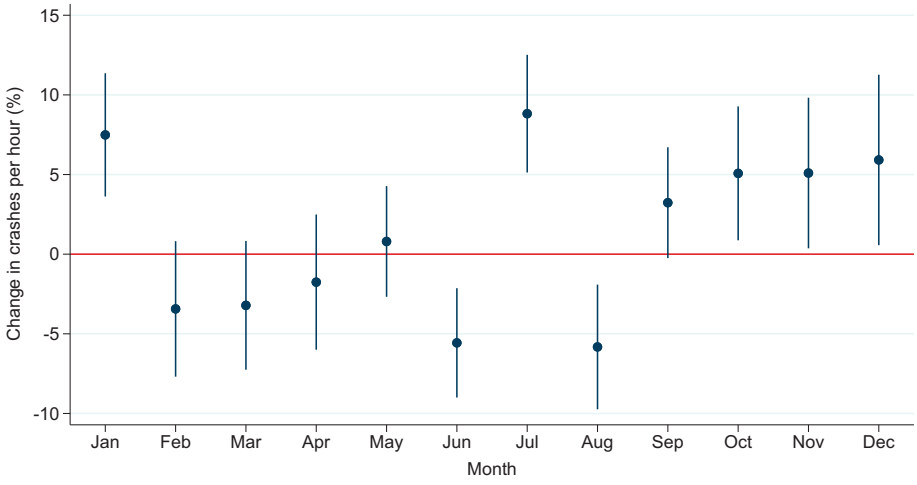


Fig. 4. Effect of fatality messages on crashes by calendar month. Shown are the δ_i coefficient estimates (circles) and the associated 95% confidence intervals (bars) from the regression below that allows the treatment effect to vary by calendar month: $\text{month}_{d,i}$ as an indicator if day d occurs during calendar month i. Remaining variables are defined in Fig. 2 (see also table S9). Standard errors are clustered by geography-year-month bins, where geography bins are defined as the 10^2 km^2 containing the DMS. The sample period is January 2010 to December 2017. The equation used was as follows:

$$\text{crash}(\%)_{s(10),d,h} = \sum_{i\in\{\text{Jan},\ldots,\text{Dec}\}} \delta_i \bullet \text{campaign week}_{d,h} \bullet \text{month}_{d,i} \bullet \text{post}_d + \sum_{i\in\{\text{Jan},\ldots,\text{Dec}\}} \beta_{1,i} \bullet \text{campaign week}_{d,h} \bullet \text{month}_{d,i} + \beta_2 \bullet \text{trace precipitation}_{s,d,h} + \beta_3 \bullet \text{trace precipitation}_{s,d,h} \bullet \text{post}_d + \beta_4 \bullet \text{precipitation}_{s,d,h} + \beta_5 \bullet \text{precipitation}_{s,d,h} \bullet \text{post}_d + \gamma_{s,m(d),dow(d),h} + \zeta_{\text{holiday}} + \varepsilon_{s,d,h}$$

that larger fatality numbers are more distracting and is inconsistent with the variation over the year simply being due to seasonal weather or driving patterns.

Fourth, the increase in crashes is larger in areas that place high cognitive loads on drivers. We used three related measures for whether a road segment is complex and would require high cognitive loads: centerline kilometers, lane kilometers, and average daily vehicle kilometers traveled (VKTs) (*28*). We normalized these measures to have a mean of 0 and a standard deviation of 1, and interacted them with our treatment variable. As columns 1 to 3 of Table 2 show, we found that all three measures of complexity are associated with more crashes during campaign weeks. The first row shows that a 1 standard deviation increase in any of our measures of complexity is associated with 2.1 to 3.1% more crashes during treated weeks. The second row shows that for road segments of average complexity, displaying a fatality message is also associated with more crashes (statistically significant when complexity was measured using centerline kilometers). The third and fourth rows show that, as expected, these measures are not associated with an increase in crashes during the week before a board meeting in the pretreatment period. Table S5 reports results using an indicator for whether each complexity measure is above or below the median, rather than using a continuous measure of complexity, and produces similar results.

Fifth, and closely related, the increase in crashes is higher on segments with nearby upstream DMSs. We measured the distance (on the road network) to the nearest upstream DMS, standardized this measure to have a mean of 0 and a standard deviation of 1, and multiplied by –1 so that the measure is increasing in proximity to an upstream DMS. As column 4 of Table 2 shows, fatality messages displayed on DMSs with an average proximity to an upstream DMS are associated with a 1.35% increase in crashes (P = 0.024), and increasing the closeness of the nearest upstream DMS by 1 standard deviation is associated with an incremental 0.6% increase in crashes (P = 0.026).

This fifth finding is consistent with three explanations. First, it is consistent with fatality messages having larger effects when drivers face high cognitive loads, because drivers have likely seen multiple DMS messages on these segments. Second, it is consistent with an effect caused by repeated exposures either because it means that more drivers have seen the message at least once, distracting multiple drivers, or because seeing a fatality message repeatedly in quick succession increases the message's salience (and cognitive load). Finally, an increase in crashes on segments with

Table 2. Effect of fatality messages on crashes: Segment characteristics. Shown are estimates of how the effect of campaign weeks on traffic crashes varies by segment characteristics. "Measure" is one of the following characteristics of segment s (as indicated in the column header) standardized to have a mean of 0 and a standard deviation of 1: Centerline km, Lane km, VKT, and DMS proximity. See Table 1 for additional details and table S9 for detailed variable definitions. Standard errors are clustered by geography-year-month and are shown in parentheses. ***$P < 0.01$; **$P < 0.05$. The equation used was as follows: $\text{crash}(\%)_{s(10),d,h} = \delta_1 \cdot \text{campaign week}_{d,h} \cdot \text{measure}_s + \text{post} + \delta_2 \cdot \text{campaign week}_{d,h} \cdot \text{post} + \beta_1 \cdot \text{campaign week}_{d,h} \cdot \text{measure}_s + \beta_2 \cdot \text{campaign week}_{d,h} + \beta_3 \cdot \text{trace precipitation}_{s,d,h} + \beta_4 \cdot \text{trace precipitation}_{s,d,h} \cdot \text{post}_d + \beta_5 \cdot \text{precipitation}_{s,d,h} + \beta_6 \cdot \text{precipitation}_{s,d,h} \cdot \text{post}_d + \gamma_{s,m(d),dow(d),h} + \zeta_{\text{holiday}} + \varepsilon_{s,d,h}$

	Crashes per hour >10 km (%)			
	Centerline km	Lane km	VKT	DMS proximity
	(1)	(2)	(3)	(4)
Campaign week × measure × post	2.05 (0.82)**	2.80 (0.98)***	3.05 (0.95)***	0.60 (0.27)**
Campaign week × post	1.61 (0.71)**	1.06 (0.69)	1.05 (0.69)	1.35 (0.60)**
Campaign week × measure	0.23 (0.53)	0.38 (0.71)	0.12 (0.67)	0.06 (0.20)
Campaign week	−0.21 (0.51)	−0.05 (0.55)	−0.04 (0.55)	−0.33 (0.43)
Observations	48,236,425	53,648,884	53,648,884	61,627,553
Adjusted R^2	0.08	0.08	0.08	0.08
Rain and interactions	Yes	Yes	Yes	Yes
S-Y-M-D-H FE	Yes	Yes	Yes	Yes
Holiday FE	Yes	Yes	Yes	Yes

Table 3. Effect of fatality messages by crash types. Shown are estimates of the effect of campaign weeks on single- and multi-vehicle crashes. The dependent variable is the number of crashes occurring over the 10 km downstream of DMS s on date d during hour h of a specific type, scaled by the population average for all segments of that type and multiplied by 100. See Table 1 for additional details. Standard errors are clustered by geography-year-month and are shown in parentheses. **$P < 0.05$. The equation used was as follows: $\text{crash}(\%)_{s(10),d,h} = \delta \cdot \text{campaign week}_{d,h} \cdot \text{post}_d + \beta_1 \cdot \text{campaign week}_{d,h} + \beta_2 \cdot \text{trace precipitation}_{s,d,h} + \beta_3 \cdot \text{trace precipitation}_{s,d,h} \cdot \text{post}_d + \beta_4 \cdot \text{precipitation}_{s,d,h} + \beta_5 \cdot \text{precipitation}_{s,d,h} \cdot \text{post}_d + \gamma_{s,m(d),dow(d),h} + \zeta_{\text{holiday}} + \varepsilon_{s,d,h}$

	Crashes per hour (%)	
	Multi-vehicle	Single-vehicle
	(1)	(2)
Campaign week × post	1.60 (0.63)**	−0.26 (1.59)
Campaign week	−0.64 (0.44)	1.74 (1.13)
Observations	61,697,666	61,697,666
Adjusted R^2	0.08	0.01
Rain and interactions	Yes	Yes
S-Y-M-D-H FE	Yes	Yes
Holiday FE	Yes	Yes

nearby upstream DMSs is also consistent with treatment mattering, because at least one of these DMSs is likely to have displayed a fatality message.

Sixth, fatality messages increase multi-vehicle crashes but not single-vehicle crashes. In Table 3, we separately examined whether multi- and single-vehicle crashes change the week before a board meeting. We found a 1.60% increase in crashes involving multiple vehicles ($P = 0.011$) and an insignificant change in crashes involving single vehicles. Because single-vehicle crashes are likely a result of large mistakes (e.g., driving off the road), the increase in multi-vehicle crashes suggests that more small driving mistakes occur when fatality messages are displayed that are plausibly related to distracted driving (e.g., drifting out of the lane). An increase in multi-vehicle crashes is also consistent with fatality messages inducing more anxiety, and thus being more distracting, when driving conditions could be perceived as more dangerous (i.e., when other vehicles are nearby).

Seventh, the concentrated effect immediately after DMSs and decreasing effect size over longer distances previously documented is consistent with a temporary distraction effect. Prior research finds that the time to resume a task after an interruption increases with the complexity of the interruption (*29*). Shocking, salient fatality messages plausibly present such interruptions, and drivers are expected to eventually regain their ability to safely respond to changes in traffic conditions. At 100 km/h (62 mph, normal highway speeds), a driver will travel 5 km in 3 min. Our evidence thus suggests that the distracting effect of fatality messages lasts for more than a trivial amount of time, but that drivers do recover.

Finally, our proposed mechanism is consistent with evidence from the traffic safety literature. Most directly related, Shealy *et al.* (*15*) found, in a laboratory setting, that showing drivers fatality messages increased neurocognition, which is a proxy for attention/working memory and cognitive load. Although it is difficult to compare estimates across experiments, a rough estimate is that showing drivers a fatality message increases cognitive load by 50% (*15*, *30*). Although there remains a debate on whether billboards cause crashes (*31*), recent studies using vehicle simulators have found that, depending on the content, billboards do cause people to drive worse as measured by variability in speed and lane position, reaction times, vehicle headway, and number of crashes (*32*–*34*). Furthermore, studies using vehicle simulators have found that increasing individuals' anxiety causes them to drive worse, and that these effects can last for at least 2 km (*35*, *36*). Thus, prior traffic safety research, combined with our seven pieces of evidence, provides strong support for a temporary distraction effect caused by fatality messages that reduces individuals' ability to drive safely and respond to changes in traffic conditions.

Alternative hypotheses

In the supplementary text, section S4, we address seven alternative hypotheses, including the possibilities that treated weeks are inherently more dangerous, that fatality messages help in the long run or result in improvements away from DMSs, that displaying any message causes crashes, and that fatality messages cause some drivers to slow down, increasing the variance of vehicle speeds and thus crash risk. We provide evidence inconsistent with each of these alternative hypotheses.

Robustness

In table S6 we report several robustness tests of our difference-in-differences estimates. In particular, we show that clustering by segment-year-month reduces the standard error in half, clustering by just geography

Table 4. Effect of fatality messages on statewide crashes. Shown are estimates of the effect of campaign weeks on statewide crashes. The dependent variable is the number of crashes occurring statewide (column 1), statewide on the highway system (column 2), or statewide off the highway system (column 3) on date d during hour h, scaled by the population average and multiplied by 100. We include year-month-day-of-week-hour and holiday fixed effects. Standard errors are clustered by year-month and are shown in parentheses. ***$P < 0.01$; **$P < 0.05$. The equation used was as follows: statewide crashes $(\%)_{d,h} = \delta \cdot \text{campaign week}_{d,h} \cdot \text{post}_d + \beta_1 \cdot \text{campaign week}_{d,h} + \gamma_{m(d),dow(d),h} + \zeta_{\text{holiday}} + \varepsilon_{d,h}$

	Total	On-highway	Off-highway
	(1)	(2)	(3)
Campaign week × post	1.98 (0.96)**	2.77 (1.19)**	1.16 (0.95)
Campaign week	−1.61 (0.72)**	−2.39 (0.89)***	−0.79 (0.75)
Observations	70,127	70,127	70,127
Adjusted R^2	0.87	0.82	0.84
Y-M-D-H FE	Yes	Yes	Yes
Holiday FE	Yes	Yes	Yes

produces slightly smaller standard errors, controlling for rain more flexibly does not affect our results, not controlling at all for rain doubles our estimated treatment effect, not controlling for holidays increases our estimate slightly, and dropping hours immediately before and after campaign weeks (i.e., hours outside of campaign weeks that sometimes display fatality messages, see fig. S6) further increases our estimate. Further, we show that the estimated treatment effect is larger when using alternative outcome measures; specifically, using an indicator variable for whether there is any crash or using the log of the number of crashes plus one. We did not use count data models (e.g., Poisson regression) because they require variation in the outcome within each fixed effect and are thus incompatible with our extensive fixed-effect structure.

All of our estimates so far have assumed that any DMS that existed during our sample existed for the entire sample. We also tested whether our results are robust to limiting the sample to the DMS months where each DMS existed. To do so, we collected information on when each DMS existed using Google Street View. For each DMS month, we either know a DMS exists, know it does not exist, or are unsure. To deal with this uncertainty over when they exist, we conducted two robustness tests. We first limited our sample to the DMS months in which we know the DMS exists, and then limited our sample to the DMS months in which the DMS might exist (i.e., we do not know that it does not exist). As expected, we found that including DMS months that lack an operational DMS attenuates our estimates, with the "must exist" sample leading to a higher point estimate than the "may exist" sample, which itself leads to a higher point estimate than our full sample.

Our main tests exploit GPS locations of both DMSs and crashes and uses an extensive fixed-effects structure. To evaluate whether a simpler approach provides similar conclusions, we evaluated the change in crashes statewide during campaign weeks. Results presented in Table 4 indicate that crashes, particularly on-highway crashes, also increase statewide during campaign weeks.

We also tested for an effect of fatality messages on several measures of crash seriousness. As shown in table S7, we found that the count of vehicles involved in crashes is 1.93% higher during campaign weeks in the treatment period ($P = 0.002$). We do not have the power to detect an effect on the number of deaths, number of fatal crashes, and an estimate of the social cost of these crashes, because only 0.58% of crashes have a fatality. The 95% confidence intervals for these estimates are large, and we cannot rule out meaningful treatment effects.

Discussion

We present evidence that fatality messages are too salient and distract drivers. Part of this evidence includes documenting heterogeneous treatment effects, with larger treatment effects when the message is plausibly more salient or when drivers' cognitive loads are higher. This same evidence suggests that there are times and places where displaying fatality messages does help. Specifically, these messages reduce the number of crashes when the number of reported fatalities is in the bottom quartile and in places where the road network complexity is at least 1 standard deviation below its respective mean. Although these benefits do not outweigh the harm done, they show that behavioral interventions can help if they are not too salient and are delivered when individuals' cognitive loads are low.

The effect of displaying a fatality message on crashes is large relative to the simplicity of the intervention. We estimate that displaying a fatality message increases the number of crashes over the next 10 km of roadway by 4.5%. Our estimates suggest that displaying these messages causes an additional 2600 crashes per year in Texas alone (see supplementary text S5 for details). Furthermore, although we are underpowered to detect an effect on fatal crashes, if we assume a similar percentage change in fatal crashes, then fatality messages cause an additional 16 fatalities per year. Using estimates from Blincoe *et al.* (*37*), these additional crashes have a total social cost of $380 million per year. To calculate an estimate of the impact of fatality messages in the United States, we scaled our estimated treatment effect by the number of licensed drivers in the 28 treated states. Doing so suggests that across the United States, displaying these messages might cause an additional 17,000 crashes and 104 fatalities per year, with a total social cost of $2.5 billion per year.

There are two sources of important variation across states in how fatality messages are implemented. First, US states differ in how frequently they show fatality messages. This matters because we found that most of the damage is done during the first few days that the message is displayed (fig. S7). This finding implies that in states where the fatality message is displayed all the time (unless there is a more important message), such as Illinois, the effects could be more benign, and in places where fatality messages are displayed 1 day per week, such as Colorado, the effect could be worse. Second, whereas the exact text of the message is consistent across states, the displayed fatality account varies. Texas, the second-largest state in the United States, displays larger fatality counts than most other states. This matters because we found that fatality messages only hurt when the displayed fatality count is large (Fig. 3). If the negative effect of the fatality message depends on the absolute number displayed, then it will not have the same negative effect in most other states. However, if the negative effect depends on the number displayed relative to a state's population, then our results are more generalizable to other states. For additional discussion of this study's external validity, see the supplementary text, section S6.

Conclusion

Our study shows that salient, generic, in-your-face safety messages delivered to drivers crowd out more pressing safety concerns, yielding immediate negative and socially undesirable outcomes. The treatment effect is larger when the reported number of deaths is larger and when road segments are more complex. Our evidence suggests that even after several years,

an intervention delivered 1 week each month still increases crashes. Further, the negative effects of these messages appear to be constrained to the immediate vicinity and time where delivered.

These findings contribute to three areas of research. Existing research on DMS safety messages finds evidence that messages about speeding, fog, or slippery roads are effective at reducing drivers' speeds (*38*–*40*), but that generic safety messages have little effect (*41*). The traffic safety literature also finds that drivers rate negatively framed threat appraisals as more effective (*42*–*44*), but that such messages can be perceived as controlling or manipulative, potentially causing people to ignore them (*45*). Shealy *et al.* (*15*) found, in a laboratory setting, that showing drivers nontraditional safety messages, including fatality messages, increases their attention and cognitive load, which they interpreted as a good thing. By contrast, we show that this can have costly consequences. We show that, contrary to drivers' and policy-makers' expectations, using fatality messages to increase awareness of the risk of driving causes additional traffic crashes.

We also contribute to the literature on risk disclosures (*46*–*49*) and the broader literature on information disclosure (*50*–*52*). Although risk disclosures are common in many markets, many tend to be generic rather than specific (e.g., "driving is dangerous" versus "sharp turn ahead"). There is concern that generic risk disclosures may be ineffective at reducing risk-taking because of their lack of specificity, but there is limited empirical evidence on their effectiveness. Our setting allows us to measure the effectiveness of a generic risk disclosure. We found that generic, yet plausibly shocking, risk disclosures can affect individual behavior.

Finally, we contribute to the literature on behavioral interventions. An important question within this literature is whether the effects of behavioral interventions persist after treatment stops. Although there are some notable exceptions (*53*, *54*), this literature typically finds little persistence (*55*). We found no evidence that fatality messages affect behavior outside of campaign weeks. Another question in this literature is whether individuals habituate to behavioral interventions, as budget considerations make it difficult to test the long-term effects (*53*). We found that drivers do not habituate to fatality messages, potentially because the number of displayed deaths is constantly updated, with the treatment effect remaining virtually unchanged 5 years after the initial implementation. This should increase confidence that the estimated short-term benefits in other studies may persist in the long term.

This evaluation of fatality messages highlights five key lessons. First, and most importantly, behavioral interventions can fail if they increase individuals' cognitive loads to the extent that they crowd out more important considerations. Thus, given that behavioral interventions are intentionally designed to be salient and seize attention, the message, delivery, and timing must be carefully designed to prevent the intervention from backfiring. Second, because individuals face cognitive constraints, a full accounting of an intervention's welfare effects should consider whether adding to participants' cognitive loads has effects outside of the targeted domain. Third, our results speak to the trade-offs of using behavioral interventions versus taxes and subsidies to address externalities such as unsafe driving, pollution, and global warming (*56*, *57*). The general lesson is that behavioral interventions targeting externalities that are largest when individuals' cognitive loads are high are likely to be less efficient than taxes or subsidies. Fourth, measuring an intervention's effect is important, even for simple interventions, because good intentions need not imply good outcomes. Finally, and most directly, fatality message campaigns increase the number of crashes, so ceasing these campaigns is a low-cost way to improve traffic safety.

Materials and methods

Data

We collected data on traffic crashes, DMS locations and messages displayed, TxDOT's board meeting schedule, weather, the Texas road network, and US federal holidays. Table S8 summarizes the January 2010 to July 2012 "pretreatment" period (i.e., the time period before TxDOT began displaying fatality messages) and the August 2012 to December 2017 "treatment" period. We collected data on 880 DMSs.

Our data on traffic crashes comes from the TxDOT Crash Records Information System and includes all reported crashes occurring on Texas roads. This dataset includes the GPS coordinates and other characteristics for each crash from 2010 to 2017.

We collected DMS location data from the TxDOT website and from lists provided by TxDOT of all DMSs in 2013 and 2015. We combined these location data and validated and corrected them using Google Maps. We corrected 18% of the DMS locations and updated the direction of travel for 26 DMSs. We dropped 175 DMSs that were portable, test DMSs, or smaller than standard. These smaller DMSs are often just able to display a few characters and used for displaying travel times or tolls. The largest DMSs that we dropped for being too small can display two lines of 12 characters, whereas standard DMSs can display three lines of 15 or 18 characters. We also dropped nine DMSs located on local roads rather than on highways. Fig. S8 plots the locations of DMSs within the entire state, and fig. S9 plots those in the Houston area. These maps show that DMSs are located primarily within urban areas, and that within urban areas, DMSs are spaced fairly evenly apart, with a median driving distance of 5.3 km between consecutive DMSs.

We collected information on when each DMS exists using Google Street View. These data are limited because the mean gap between the last time a DMS is known not to exist and the first time it is known to exist is 2.9 years, whereas the mean gap between the last time a DMS is known to exist and the first time it is known to not exist is 1.4 years. From these data, we know that at least 24% of DMSs did not exist over our entire sample. Our main results assume that all DMSs that exist during our sample exist for the entire sample. Including nonoperational DMSs biases our results toward 0.

We gathered data on the messages displayed on DMSs from two sources. First, we obtained log files for the DMSs located in the Houston area from Houston TranStar for the years 2012–2013, and second, we collected hourly DMS message content directly from the TxDOT website for all Texas DMSs for 2016–2017.

We collected TxDOT's board meeting schedule from the TxDOT website. These meetings are typically held the last Thursday of each month, except in November and December, when they are held earlier to avoid conflicting with Thanksgiving and Christmas.

We obtained hourly weather data from the US National Oceanic and Atmospheric Administration's Integrated Surface Database. Figures S8 and S9 also show the locations of the weather stations that we used. The median distance between a DMS and the nearest weather station is 14 km.

Variable definitions

Our primary outcome variable is the hourly number of crashes on a given road segment. Road segments begin at DMS locations and continued for x km of highway driving distance, with $x \in \{-10,-9,\ldots,9,10\}$; negative distances denote segments preceding the DMS (i.e., upstream) and positive distances denote segments continuing past the DMS (i.e., downstream). We calculated driving distances using the Open Source Routing Machine and Open Street Maps data for the Texas highway network. Our network includes all roads classified as motorways, motorway links, trunk roads, and primary roads. This is the smallest set of classifications that includes all highways but also includes some roads that are not highways. Figure S10 depicts segments of 1, 3, 5, and 10 km downstream of a sample DMS near Aledo, Texas, and the crashes associated with each segment.

Because we allowed road segments to merge and diverge onto other highways, road segments

of length x + 1 km typically contain more than an additional 1 km of road surface area and thus have a more than proportional increase in the number of crashes. We therefore scaled hourly crash counts for segments of length x by the average number of crashes occurring over all segments of length x over the entire sample period to create a standardized measure of crashes that is easier to interpret. We label this variable crash(%)$_{s(x),d,h}$, where the subscripts index segment (s), segment length (x), day (d), and hour (h). See table S9 for detailed variable definitions.

Lane kilometers and average daily VKT were measured from the Highway Performance Monitoring System annually, and centerline kilometers were measured using Open Street Maps. All three were measured over each segment.

We defined campaign weeks using the schedule of TxDOT board meetings and our DMS log files. Since August 2012, TxDOT traffic engineers have been instructed to display the fatality message beginning "after morning peak" on the Monday 1 week before a board meeting and ending "before morning peak" on the following Monday. Exact times are not provided, because "morning peak" varies by highway and direction of travel.

We determined the typical start and end times of campaign weeks using our DMS log files. Figure S6 shows that a fatality message is displayed for ~8% of the DMS hours between midnight and 7:00 am on the Monday 1 week before board meetings (designated first day), increasing to 12, 18, and 29% during the 7:00, 8:00, and 9:00 a.m. hours, respectively, and is then displayed for ~29 to 43% of DMS hours during the campaign week. Percentages <100% are consistent with instructions that fatality messages "should not pre-empt needed traffic messages, incident-related messages, Emergency Operation Center messages (EOC), or Amber/Silver/Blue alerts." Fatality messages also gradually disappear at the end of the campaign week, with the fatality message displaying for 21, 14, and 11% of DMS hours during the 6:00, 7:00, and 8:00 a.m. hours of the final Monday, respectively. Thus, although there is leakage into hours immediately before and after the intended display period, we found that fatality messages concentrate during the designated week. On the basis of the patterns observed in fig. S6, we defined an indicator variable for the week before a board meeting, campaign week$_{d,h}$, which equals 1 for all days (d) and hours (h) between 9:00 a.m. on the Monday 10 days before a scheduled board meeting and 7:00 a.m. the following Monday. In the supplementary text, section S7, we further explore how campaign weeks affect the messages displayed.

To control for variation in weather conditions, we defined two indicators for whether the weather station closest to DMS s reported precipitation during hour h of day d. Specifically, trace precipitation$_{s,d,h}$ was set equal to 1 if the weather station reported <1 mm of precipitation, and 0 otherwise, and precipitation$_{s,d,h}$ was set equal to 1 if the weather station reported ≥1 mm of precipitation, and 0 otherwise.

Because we did not observe the displayed death count in every month, we imputed the year-to-date fatality count for each month using the actual number of year-to-date fatalities. From the DMS log files, we found that the reported fatality number is reported with a median lag of 22 days, and we used this lag when imputing the number of fatalities for each month.

Table S10 reports summary statistics for our data. As discussed earlier, because of the increasing surface area covered by segments of larger lengths, the number of crashes per hour increases more than proportionally in segment length, with 8.2 times more crashes within 10 km of a DMS than within 3 km. Crashes are proportional to lane kilometers and VKT (table S11).

Research design

To estimate the effect of fatality messages on the number of traffic crashes, we exploited within-month variation of fatality messages while controlling for weather, holidays, and idiosyncratic segment characteristics. To control for unobservable within-month fixed-segment characteristics (e.g., idiosyncratic elements of the season, time of day, and day of the week specific to each DMS highway segment), we included segment-year-month-day-of-week-hour fixed effects. Because fatality messages are only instructed to be displayed for 1 week each month, we could compare, for each DMS highway segment, year-month-day-of-week-hours when the message was instructed versus not instructed to be displayed. We also included controls for precipitation and holiday fixed effects. We estimated the following ordinary least-squares regression using all observations from 1 August 2012 through 31 December 2017 as follows:

$$\text{crash}(\%)_{s(x),d,h} = \delta \bullet \text{campaign week}_{d,h} + \beta_1 \bullet \text{trace precipitation}_{s,d,h} + \beta_2 \bullet \text{precipitation}_{s,d,h} + \gamma_{s,m(d),dow(d),h} + \zeta_{\text{holiday}} + \varepsilon_{s,d,h} \quad (1)$$

In Regression 1, δ is our estimated treatment effect, γ is a fixed effect for each segment-year-month-day-of-week-hour, ζ is a fixed effect for each holiday, and dow(d) is the day of the week associated with day d.

We also estimated the treatment effect using a difference-in-differences specification that exploits both within-month variation in when fatality messages are instructed to be displayed and differences between the treatment and pretreatment periods. This approach directly addresses the concern that campaign weeks could systematically differ from other weeks (e.g., total traffic volume or crash risk). Specifically, we estimated the following regression:

$$\text{crash}(\%)_{s(x),d,h} = \delta \bullet \text{campaign week}_{d,h} \bullet \text{post}_d + \beta_1 \bullet \text{campaign week}_{d,h} + \beta_2 \bullet \text{trace precipitation}_{s,d,h} + \beta_3 \bullet \text{trace precipitation}_{s,d,h} \bullet \text{post}_d + \beta_4 \bullet \text{precipitation}_{s,d,h} + \beta_5 \bullet \text{precipitation}_{s,d,h} \bullet \text{post}_d + \gamma_{s,m(d),dow(d),h} + \zeta_{\text{holiday}} + \varepsilon_{s,d,h} \quad (2)$$

which is equivalent to taking the difference between δ from Regression 1 for the August 2012 to December 2017 sample and δ from the same regression for the January 2010 to July 2012 sample. In Regression 2, δ is the coefficient of interest. In our analyses, we found no difference during the pretreatment period in downstream crashes between the week before a board meeting ("campaign weeks") and other weeks, so the primary difference between Regressions 1 and 2 is that the second has larger standard errors. This occurs because Regression 1 presumes that campaign weeks would be exactly the same as other weeks in the absence of treatment, whereas Regression 2 acknowledges uncertainty about whether campaign weeks would be the same in the post-period in the absence of treatment.

As a conservative approach, we clustered standard errors by geography-year-month, where geography refers to bins of size x^2 km^2 that contain a DMS segment of length x. Thus, fewer clusters (geographic bins larger in area) are used for segments of greater length, because crashes occurring over those longer lengths may be linked to multiple DMSs.

See the supplementary text, section S8, for additional discussion of our materials and methods.

REFERENCES AND NOTES

1. Z. Afif, W. W. Islan, O. Calvo-Gonzalez, A. Dalton, "Behavioral science around the world: Profiles of 10 countries" (World Bank Group, 2019); http://documents.worldbank.org/curated/en/710771543609067500/Behavioral-Science-Around-the-World-Profiles-of-10-Countries.
2. Organisation for Economic Co-Operation and Development, "Behavioural insights" (OECD, 2018); https://www.oecd.org/gov/regulatory-policy/behavioural-insights.htm.
3. D. P. Byrne, A. L. Nauze, L. A. Martin, Tell me something I don't already know: Informedness and the impact of information programs. *Rev. Econ. Stat.* **100**, 510–527 (2018). doi: 10.1162/rest_a_00695
4. Organisation for Economic Co-Operation and Development, "Tools and Ethics for Applied Behavioural Insights: The BASIC Toolkit" (OECD, 2019); doi: http://dx.doi.org/10.1787/9ea76a8f-en.
5. Taylor and Thompson (6) define salience as, "the phenomenon that when one's attention is differentially directed to one portion on the environment rather than to others, the information contained in that portion will receive disproportionate weighing in subsequent judgments." In the case of the intervention that we are studying, drivers' attention is being directed to the fact that many people have died while driving and away from the act of driving, leaving insufficient working memory resources to process driving conditions [see Baddeley (7)].
6. S. E. Taylor, S. C. Thompson, Stalking the elusive "vividness" effect. *Psychol. Rev.* **89**, 155–181 (1982). doi: 10.1037/0033-295X.89.2.155

7. A. Baddeley, Working memory. *Curr. Biol.* **20**, R136–R140 (2010). doi: 10.1016/j.cub.2009.12.014; pmid: 20178752
8. Centers for Disease Control and Prevention, "WISQARS™ – Web-based injury statistics query and reporting system" (CDC, 2018); https://www.cdc.gov/injury/wisqars/index.html.
9. World Health Organization, "Global health estimates 2016: Deaths by cause, age, sex, by country and by region, 2000–2016" (WHO, 2018); https://web.archive.org/web/20200421084616/https://www.who.int/healthinfo/global_burden_disease/GHE2016_Deaths_Global_2000_2016.xls.
10. K. Cunningham, S. Stratford, "ODOT to display amount of traffic deaths on digital boards along Ohio highways," *Fox 8 Cleveland*, 1 July 2015; https://fox8.com/news/odot-to-display-amount-of-traffic-deaths-on-digital-boards-along-ohio-highways/.
11. "TxDOT signs to regularly display traffic death numbers," *CBS News Dallas-Ft. Worth*, 21 August 2012; http://web.archive.org/web/20120827032750/https://dfw.cbslocal.com/2012/08/21/txdot-signs-to-regularly-display-traffic-death-numbers/.
12. See table S1 for a list of states that have displayed a fatality message. Fatality messages have also been used in at least one other country (South Korea).
13. J. Brandel, "What's the deal with Illinois' traffic death highway signs?," *WBEZ Chicago*, 9 April 2013; https://web.archive.org/web/20200930222330/https://www.wbez.org/stories/whats-the-deal-with-illinois-traffic-death-highway-signs/f38eaf02-f513-4898-9875-70fc425e54ab.
14. L. Boyle, G. Cordahi, K. Grabenstein, M. Madi, E. E. Miller, P. Silberman, "Effectiveness of safety and public service announcement messages on dynamic message signs" (Federal Highway Administration, Technical Report FHWA-HOP-14-015, 2014); https://ops.fhwa.dot.gov/publications/fhwahop14015/technical.htm.
15. T. Shealy, P. Kryschtal, K. Franczek, B. J. Katz, "Driver response to dynamic message sign safety campaign messages" (Virginia Transportation Research Council, Technical Report VTRC 20-R16, 2020); https://rosap.ntl.bts.gov/view/dot/54604.
16. We do not compare hours in which a fatality message is displaying versus not displaying because whether a fatality message is displaying is endogenous. We provide evidence of this endogeneity in the supplementry text, section S1.
17. A. van Benthem, What is the optimal speed limit on freeways? *J. Public Econ.* **124**, 44–62 (2015). doi: 10.1016/j.jpubeco.2015.02.001
18. G. DeAngelo, B. Hansen, Life and death in the fast lane: Police enforcement and traffic fatalities. *Am. Econ. J. Econ. Policy* **6**, 231–257 (2014). doi: 10.1257/pol.6.2.231
19. We note that the point estimates in Fig. 2 increase over time; one possible explanation for this trend is a decline in leakage of the fatality message outside of the designated campaign weeks.
20. D. Kahneman, *Attention and Effort* (Prentice-Hall, 1973).
21. J. Sweller, J. J. van Merriënboer, F. Paas, Cognitive architecture and instructional design: 20 years later. *Educ. Psychol. Rev.* **31**, 261–292 (2019). doi: 10.1007/s10648-019-09465-5
22. J. W. de Fockert, G. Rees, C. D. Frith, N. Lavie, The role of working memory in visual selective attention. *Science* **291**, 1803–1806 (2001). doi: 10.1126/science.1056496; pmid: 11230699
23. D. T. Gilbert, B. W. Pelham, D. S. Krull, On cognitive busyness: When person perceivers meet persons perceived. *J. Pers. Soc. Psychol.* **54**, 733–740 (1988). doi: 10.1037/0022-3514.54.5.733
24. N. Berggren, S. B. Hutton, N. Derakshan, The effects of self-report cognitive failures and cognitive load on antisaccade performance. *Front. Psychol.* **2**, 280 (2011). doi: 10.3389/fpsyg.2011.00280; pmid: 22046166
25. D. L. Strayer, J. M. Cooper, J. Coleman, N. Medeiros-Ward, F. Biondi, "Measuring cognitive distraction in the automobile" (AAA Foundation for Traffic Safety, 2013); https://aaafoundation.org/wp-content/uploads/2018/01/MeasuringCognitiveDistractionsReport.pdf.
26. R. M. Yerkes, J. D. Dodson, The relation of strength of stimulus to rapidity of habit-formation. *J. Comp. Neurol. Psychol.* **18**, 459–482 (1908). doi: 10.1002/cne.920180503
27. M. A. Staal, "Stress, cognition, and human performance: A literature review and conceptual framework" (NASA Technical Report TM-2004-212824, 2004); https://humansystems.arc.nasa.gov/publications/20051028105746_IH-054 Staal.pdf.
28. Consider the following three examples of a 10-km segment to contrast centerline and lane kilometers: (i) a straight road with four lanes; (ii) a Y-shaped road that splits halfway, with all parts having four lanes; and (iii) a Y-shaped road that splits halfway where the trunk is four lanes and each branch is two lanes. Segment (i) has 10 centerline km and 40 lane km, segment (ii) has 15 centerline km and 60 lane km, and (iii) has 15 centerline km and 40 lane km.
29. C. A. Monk, D. A. Boehm-Davis, J. G. Trafton, Recovering from interruptions: Implications for driver distraction research. *Hum. Factors* **46**, 650–663 (2004). doi: 10.1518/hfes.46.4.650.56816; pmid: 15709327
30. H. Tsunashima, K. Yanagisawa, Measurement of brain function of car driver using functional near-infrared spectroscopy (fNIRS). *Comput. Intell. Neurosci.* **2009**, 164958 (2009). doi: 10.1155/2009/164958; pmid: 19584938
31. O. Oviedo-Trespalacios, V. Truelove, B. Watson, J. A. Hinton, The impact of road advertising signs on driver behaviour and implications for road safety: A critical systematic review. *Transp. Res. Part A Policy Pract.* **122**, 85–98 (2019). doi: 10.1016/j.tra.2019.01.012
32. K. Mollu, J. Cornu, K. Brijs, A. Pirdavani, T. Brijs, Driving simulator study on the influence of digital illuminated billboards near pedestrian crossings. *Transp. Res., Part F Traffic Psychol. Behav.* **59**, 45–56 (2018). doi: 10.1016/j.trf.2018.08.013
33. H. Marciano, The effect of billboard design specifications on driving: A driving simulator study. *Accid. Anal. Prev.* **138**, 105479 (2020). doi: 10.1016/j.aap.2020.105479; pmid: 32178794
34. L. Meuleners, P. Roberts, M. Fraser, Identifying the distracting aspects of electronic advertising billboards: A driving simulation study. *Accid. Anal. Prev.* **145**, 105710 (2020). doi: 10.1016/j.aap.2020.105710; pmid: 32777558
35. G. F. Briggs, G. J. Hole, M. F. Land, Emotionally involving telephone conversations lead to driver error and visual tunnelling. *Transp. Res., Part F Traffic Psychol. Behav.* **14**, 313–323 (2011). doi: 10.1016/j.trf.2011.02.004
36. E. Roidl, B. Frehse, R. Höger, Emotional states of drivers and the impact on speed, acceleration and traffic violations - a simulator study. *Accid. Anal. Prev.* **70**, 282–292 (2014). doi: 10.1016/j.aap.2014.04.010; pmid: 24836476
37. L. Blincoe, T. R. Miller, E. Zaloshnja, B. A. Lawrence, "The economic and social impact of motor vehicle crashes, 2010 (revised)" (National Highway Traffic Safety Administration, Technical Report DOT HS 812 013, 2015); http://www-nrd.nhtsa.dot.gov/Pubs/812013.pdf.
38. P. Rämä, R. Kulmala, Effects of variable message signs for slippery road conditions on driving speed and headways. *Transp. Res., Part F Traffic Psychol. Behav.* **3**, 85–94 (2000). doi: 10.1016/S1369-8478(00)00018-8
39. A. S. Al-Ghamdi, Experimental evaluation of fog warning system. *Accid. Anal. Prev.* **39**, 1065–1072 (2007). doi: 10.1016/j.aap.2005.05.007; pmid: 17920827
40. N. Chaurand, F. Bossart, P. Delhomme, A naturalistic study of the impact of message framing on highway speeding. *Transp. Res., Part F Traffic Psychol. Behav.* **35**, 37–44 (2015). doi: 10.1016/j.trf.2015.09.001
41. A. H. Jamson, N. Merat, "The effectiveness of safety campaign VMS messages - A driving simulator investigation," in *Proceedings of the 4th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design, Washington, July 9–12, 2007* (University of Iowa, 2007); pp. 459–465. doi: https://doi.org/10.17077/drivingassessment.1277
42. R. L. Cathcart, A. I. Glendon, Judged effectiveness of threat and coping appraisal anti-speeding messages. *Accid. Anal. Prev.* **96**, 237–248 (2016). doi: 10.1016/j.aap.2016.08.005; pmid: 27544888
43. S.-A. Kaye, M. J. White, I. M. Lewis, Individual differences in drivers' cognitive processing of road safety messages. *Accid. Anal. Prev.* **50**, 272–281 (2013). doi: 10.1016/j.aap.2012.04.018; pmid: 22608267
44. C. G. Sibley, N. Harré, The impact of different styles of traffic safety advertisement on young drivers' explicit and implicit self-enhancement biases. *Transp. Res., Part F Traffic Psychol. Behav.* **12**, 159–167 (2009). doi: 10.1016/j.trf.2008.11.001
45. N. J. Ward, K. Finley, A. Townsend, B. G. Scott, The effects of message threat on psychological reactance to traffic safety messaging. *Transp. Res., Part F Traffic Psychol. Behav.* **80**, 250–259 (2021). doi: 10.1016/j.trf.2021.04.013
46. R. Forsythe, R. Lundholm, T. Rietz, Cheap talk, fraud, and adverse selection in financial markets: Some experimental evidence. *Rev. Financ. Stud.* **12**, 481–518 (1999). doi: 10.1093/revfin/12.3.0481
47. G. Z. Jin, P. Leslie, The effect of information on product quality: Evidence from restaurant hygiene grade cards. *Q. J. Econ.* **118**, 409–451 (2003). doi: 10.1162/003355303321675428
48. D. R. Longo, Risk, communication and health psychology. *Health Expect.* **8**, 186–187 (2005). doi: 10.1111/j.1369-7625.2005.00328.x
49. J. M. Madsen, J. L. McMullin, Economic consequences of risk disclosures: Evidence from crowdfunding. *Account. Rev.* **95**, 331–363 (2020). doi: 10.2308/accr-52641
50. D. Dranove, G. Z. Jin, Quality disclosure and certification: Theory and practice. *J. Econ. Lit.* **48**, 935–963 (2010). doi: 10.1257/jel.48.4.935
51. S. Kapoor, A. Magesan, Paging Inspector Sands: The costs of public information. *Am. Econ. J. Econ. Policy* **6**, 92–113 (2014). doi: 10.1257/pol.6.1.92
52. P. Adams, S. Hunt, C. Palmer, R. Zaliauskas, Testing the effectiveness of consumer financial disclosure: Experimental evidence from savings accounts. *J. Financ. Econ.* **141**, 122–147 (2021). doi: 10.1016/j.jfineco.2020.05.009
53. H. Allcott, T. Rogers, The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *Am. Econ. Rev.* **104**, 3003–3037 (2014). doi: 10.1257/aer.104.10.3003
54. M. Bernedo, P. J. Ferraro, M. Price, The persistent impacts of norm-based messaging and their implications for water conservation. *J. Consum. Policy.* **37**, 437–452 (2014). doi: 10.1007/s10603-014-9266-0
55. A. Brandon, P. Ferraro, J. List, R. Metcalfe, M. Price, F. Rundhammer, "Do the effects of social nudges persist? Theory and evidence from 38 natural field experiments" (ExCEN, Working Paper 23277, 2017); https://dx.doi.org/10.3386/w23277.
56. E. Farhi, X. Gabaix, Optimal taxation with behavioral agents. *Am. Econ. Rev.* **110**, 298–336 (2020). doi: 10.1257/aer.20151079
57. B. D. Bernheim, D. Taubinsky, "Behavioral public economics" (National Bureau of Economic Research, Working Paper 24828, 2018); https://www.nber.org/papers/w24828.
58. J. D. Hall, J. Madsen, Data for: Can behavioral interventions be too salient? Evidence from traffic safety messages, Scholars Portal Dataverse (2022); doi: 10.5683/SP3/MCH1EF

ACKNOWLEDGMENTS

We thank H. Allcott, J. Campos, B. Enke, E. Glaeser, B. Greenwood, S. Heblich, M. Jacobsen, S. Jacobson, M. Kearney, K. Kroft, L. Martin, R. McMillan, P. Morrow, K. Muralidharan, C. Palsson, S. Peltzman, D. Pope, M. Shaver, D. Taubinsky, K. Vohs, and J. Waldfogel for helpful feedback; B. Stanford and E. Oeding of the TxDOT and C. Allen of Houston TranStar for help acquiring data and institutional details; and M.-A. Schmidt, M. Killian, J. Cairncross, and J. Allen for excellent research support. **Funding:** This work was supported by the Social Sciences and Humanities Research Council of Canada (grant no. 430-2015-00727) and the European Union's Horizon 2020 research and innovation programme (grant no. 101022491). The first grant, awarded in 2015, serves as an informal preregistration for this project. **Author contributions:** Both authors contributed equally to this project. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data and code are available at Scholars Portal Dataverse (*58*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abm3427
Supplementary Text S1 to S8
Figs. S1 to S18
Tables S1 to S19
References (*59*–*72*)

10 September 2021; accepted 3 March 2022
10.1126/science.abm3427

Your Legacy to Science

AN ESTATE GIFT TO THE
AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

Since 1848, our founding year, the American Association for the Advancement of Science (AAAS) has been deeply committed to advancing science, engineering and innovation around the world for the benefit of all people.

By making AAAS a beneficiary of your will, trust, retirement plan or life insurance policy, you become a member of our 1848 Society, joining Thomas Edison, Alexander Graham Bell and the many distinguished individuals whose vision led to the creation of AAAS and our world-renowned journal, *Science*, so many years ago.

Unlike many of its peers, *Science* is not for-profit. Your estate gift would provide long-term financial stability and durable annual income that will support operations and competitive innovation for years to come. **This support is vital.**

"As a teacher and instructor, I bear responsibility for the younger generations. If you have extra resources, concentrate them on organizations, like AAAS, that are doing work for all."

—Prof. Elisabeth Ervin-Blankenheim, 1848 Society member

If you intend to include AAAS in your estate plans, provide this information to your lawyer or financial adviser:

Legal Name: American Association for the Advancement of Science
Federal Tax ID Number: 53-0196568
Address: 1200 New York Avenue, NW, Washington, DC 20005

If you would like more information on making an estate gift to AAAS, cut out and return the form below or send an email to philanthropy@aaas.org. Additional details are also available online at **www.aaas.org/1848Society**.

1848 SOCIETY
AAAS

cut here

Yes, I would like more information about joining the AAAS 1848 Society.

PLEASE CONTACT ME AT:

Name: _____

Address: _____

City: _____ State: _____ Zip code: _____ Country: _____

Email: _____ Phone: _____

RETURN THIS FORM TO:
AAAS Office of Philanthropy and Strategic Partnerships • 1200 New York Avenue, NW • Washington, DC 20005 USA

RESEARCH ARTICLES

ELECTROOPTICS

Ferroelectric crystals with giant electro-optic property enabling ultracompact Q-switches

Xin Liu[1]†, Peng Tan[2,3]†, Xue Ma[4], Danyang Wang[5], Xinyu Jin[2], Yao Liu[1], Bin Xu[4], Liao Qiao[1], Chaorui Qiu[1], Bo Wang[6], Weigang Zhao[1], Chaojie Wei[7], Kexin Song[1], Haisheng Guo[1], Xudong Li[7], Sean Li[5], Xiaoyong Wei[1], Long-Qing Chen[6], Zhuo Xu[1], Fei Li[1]*, Hao Tian[2]*, Shujun Zhang[8]*

Relaxor-lead titanate ($PbTiO_3$) crystals, which exhibit extremely high piezoelectricity, are believed to possess high electro-optic (EO) coefficients. However, the optical transparency of relaxor-$PbTiO_3$ crystals is severely reduced as a result of light scattering and reflection by domain walls, limiting electro-optic applications. Through synergistic design of a ferroelectric phase, crystal orientation, and poling technique, we successfully removed all light-scattering domain walls and achieved an extremely high transmittance of 99.6% in antireflection film–coated crystals, with an ultrahigh EO coefficient r_{33} of 900 picometers per volt (pm V^{-1}), >30 times as high as that of conventionally used EO crystals. Using these crystals, we fabricated ultracompact EO Q-switches that require very low driving voltages, with superior performance to that of commercial Q-switches. Development of these materials is important for the portability and low driving voltage of EO devices.

Precise control of light propagation and intensity is critical to numerous photonic devices ranging from lasers to optical amplifiers and modulators (*1*, *2*). The rapid and efficient control of optical signals through electrical stimulus requires electro-optic (EO) materials that exhibit large changes in their refractive indices n in response to an applied electric field; e.g., the Pockels effect (*3*–*5*). The primary merits of the Pockels effect include the linear correlation between the change in refractive indices n and the applied electric field, fast responses, and a strong ability to control light, all of which are central to a wide range of photonics applications.

Ferroelectric crystals represented by $LiNbO_3$ (LN) and KD_2PO_4 (DKDP) are an important component of existing EO devices because of their availability in large size and good temperature stability (*6*, *7*). DKDP has very high optical damage thresholds suited for high-power Q-switch use (*8*). However, its hygroscopic feature requires careful protection against moisture, and thus DKDP based on EO devices require complicated fabrication processes (*9*). Further, the relatively low effective EO coefficients r_c of the LN and DKDP crystals (~21 and 24 pm V^{-1}, respectively) require the use of high voltage and/or thick material in EO devices, leading to high auxiliary cost (high voltage power supply) and difficulty in miniaturization (*10*). This issue has become a key obstacle for improving device performance. Therefore, the discovery and employment of alternative materials possessing larger Pockels effects to minimize the driving voltage and size of EO devices is highly desirable.

Many perovskite ferroelectric single crystals with substantial Pockels coefficients on the order of 10^2 pm V^{-1} —such as $BaTiO_3$, $KNbO_3$, and $Pb(Mg_{1/3}Nb_{2/3})O_3$–$PbTiO_3$ (PMN-PT)—have been developed to address the issues (*11*–*13*). Despite the promising EO properties, optical clarity of perovskite ferroelectrics is a long-standing challenge that greatly hinders practical applications of these crystals. The presence of naturally occurring ferroelectric domain walls in the crystals drastically limits their optical transparency as a result of light scattering and reflection at domain walls arising from the difference in refractive indices of the adjacent domains with different orientations, rendering high optical loss or even opacity in the visible-to-near-infrared spectrum. In the past few decades, considerable efforts have been made to eliminate the light-scattering domain walls in these ferroelectrics, but with very limited success (*13*–*21*). For example, poling a ferroelectric crystal along its polar direction can achieve a single-domain state without domain walls. However, the EO coefficients of single-domain crystals are much smaller than those of the crystals poled along a specific nonpolar direction, i.e., a domain-engineered state (*13*). More recent efforts to manipulate the domain structures of PMN-PT single crystals through ac electric field poling has demonstrated a viable approach to largely reduce the light-scattering domain walls in domain-engineered PMN-PT crystals (*19*). Although the optical transparency along the poling direction is greatly improved and a relatively high EO coefficient (r_{33}~220 pm V^{-1}) can be achieved on the basis of this method, the crystal remains opaque along the other orthogonal directions (fig. S1). Thus, designing EO devices on the basis of PMN-PT crystals remains difficult because of the challenges in obtaining high optical clarity in conjunction with giant EO performance in the crystals.

We developed a specific high-temperature poling process for [011]-oriented $Pb(In_{1/2}Nb_{1/2})O_3$-$Pb(Mg_{1/3}Nb_{2/3})O_3$-$PbTiO_3$ (PIN-PMN-PT) relaxor ferroelectric crystals to boost transparency in mutually orthogonal directions through removal of undesired domain walls, and to achieve high EO coefficients through the facilitated electric field–induced polarization rotation. This highly transparent PIN-PMN-PT crystal exhibits an ultrahigh EO coefficient r_{33} in the range of 900 to 2800 pm V^{-1}, with a temperature range of 20° to 100°C and a frequency range of 10 to 10^4 Hz. We used such poled PIN-PMN-PT crystals to construct an ultracompact free-space EO Q-switch and demonstrated its feasibility and effectiveness in miniaturization and driving voltage reduction as compared with the state-of-the-art EO devices.

Selection of relaxor ferroelectric crystals

We selected the rhombohedral 0.21PIN-(0.79 – x)PMN-xPT (x = 0.28, 0.30, 0.32) crystals as example materials, as they possess comparable room-temperature electrical properties (e.g., dielectric and piezoelectric properties) but have much improved temperature and electric field stabilities compared with those of the actively studied relaxor ferroelectric crystals PMN-PT (*22*). To achieve high transparency along both the poling and transverse directions in PIN-PMN-PT crystals, we selected the rhombohedral PIN-PMN-PT crystal with three principal axes along the crystallographic [0$\bar{1}$1], [100], and [011] directions. We show phase-field simulated domain structures for the rhombohedral crystal poled along the [011]direction (Fig. 1, A to C). Only two domain variants with polarization along the [111] and [$\bar{1}$11] directions remain in the crystal, forming 71° domain walls. The horizontal 71° domain walls are curved in the view of the (100)plane, whereas they are almost perfectly straight (parallel along the [100] direction) in the view of the (0$\bar{1}$1)

[1]Electronic Materials Research Lab, Key Lab of Education Ministry and State Key Laboratory for Mechanical Behavior of Materials, School of Electronic Science and Engineering, Xi'an Jiaotong University, Xi'an, 710049, China. [2]School of Physics, Harbin Institute of Technology, Harbin 150001, China. [3]Center of Ultra-Precision Optoelectronic Instrument Engineering, Key Laboratory of Micro-Systems and Micro-Structures Manufacturing of Ministry of Education, Harbin Institute of Technology, Harbin 150001, China. [4]School of Physical Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China. [5]School of Materials Science and Engineering, The University of New South Wales, Sydney, NSW 2052, Australia. [6]Department of Materials Science and Engineering, Materials Research Institute, The Pennsylvania State University, University Park, PA 16802, USA. [7]National Key Laboratory of Science and Technology on Tunable Laser, Harbin Institute of Technology, Harbin 150001, China. [8]Institute for Superconducting and Electronic Materials, Australian Institute for Innovative Materials, University of Wollongong, Wollongong, NSW 2500, Australia.
***Corresponding author. Email: ful5@xjtu.edu.cn (F.L.); tianhao@hit.edu.cn (H.T.); shujun@uow.edu.au (S.Z.)**
†These authors contributed equally to this work.

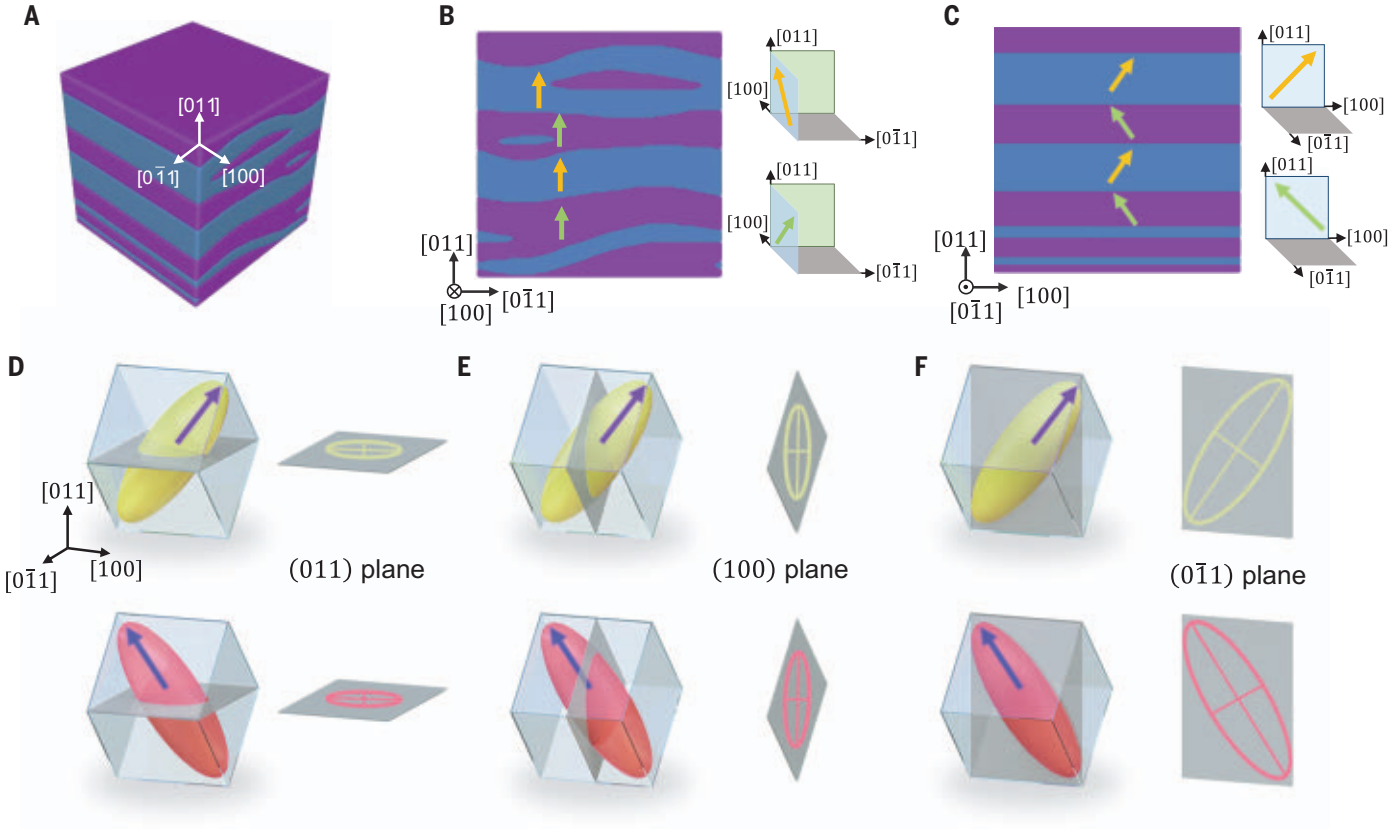


Fig. 1. Simulated domain patterns and optical indicatrix of each domain in a [011]-poled rhombohedral PIN-PMN-PT crystal. (**A**) Domain structures in the [011]-poled PIN-PMN-PT crystal under stress-free conditions on the basis of phase field simulations. (**B**) Domain pattern on a (100)plane. (**C**) Domain pattern on a (0$\bar{1}$1) plane. Purple and navy colors represent different ferroelectric domains with polarization vectors along the [$\bar{1}$11] (green) and [111](yellow) directions, respectively. (**D** to **F**) Schematic diagrams of optical indicatrix for both domains projected on the (011), (100), and (0$\bar{1}$1) planes, respectively. As shown here, the principal axes of the refractive indicatrix in both the [111] and [$\bar{1}$11] domains projected on the (011) or (100) planes are the same, leading to suppressed light scattering and/or reflection along the [011] and [100] axes. By contrast, projections of the principal axis of the optical indicatrix in neighboring domains on the (0$\bar{1}$1) plane form an angle of 71°. The alternation of the refractive indices n_o and n_e causes scattering and/or reflection of light when light travels along the [0$\bar{1}$1] direction.

plane (Fig. 1, B and C). This feature is due to the fact that the rotation of the 71° domain wall plane around the [0$\bar{1}$1] direction substantially increases the electrical and elastic energies compared with that rotates around the [100] direction (fig. S2). Therefore, the fluctuation of the 71° domain walls in the (0$\bar{1}$1) plane is prohibited, whereas small fluctuation is allowed in the (100) plane.

Considering the characteristics of the refractive indices in the [011]-poled rhombohedral crystal, high optical transparency is expected along the [011] or [100] direction. Projections of the principal axes of the optical indicatrix for the [111] and [$\bar{1}$11] domains on the (011) plane are identical (Fig. 1D). The same scenario holds true for the projections on the (100) plane (Fig. 1E). This observation suggests that the refractive indices remain unchanged as light goes across the 71° domain walls for the [011] or [100] crystallographic direction, resulting in the suppression of light scattering and/or reflection. By contrast, the optical transparency is low along the [0$\bar{1}$1] direction, considering that projection of the principal axes of the optical indicatrix on the (0$\bar{1}$1) plane is different for the [111] and [$\bar{1}$11] domains (Fig. 1F).

In addition to optical transparency, the [011]-poled rhombohedral crystals are also expected to possess high electro-optic coefficients. As the electric field is applied along the [011] direction—which is not parallel to the polar directions of the two domains, i.e., [111] or [$\bar{1}$11]—the polarization rotation in the domains can be induced (*22*, *23*). One of the most important features of relaxor ferroelectric crystals such as PIN-PMN-PT is the ease of polarization rotation under an external electric field along the nonpolar direction, which is associated with the flattened free energy landscape near the morphotropic phase boundary and the presence of nanoscale local structure heterogeneity (*23*, *24*). Consequently, the rotation of optical indicatrix in relaxor ferroelectric crystals may occur more readily in contrast to the classical ferroelectrics when subjected to external stimuli such as an electric field, leading to larger electro-optic activity in relaxor ferroelectrics.

Poling of the PIN-PMN-PT crystal

Ideally, only two kinds of ferroelectric domains (i.e., the domains with polarization along the [111] and [$\bar{1}$11] directions) should be present in the [011]-poled rhombohedral PIN-PMN-PT crystal (*25*) (Fig. 1A). In reality, however, completely removing the four electric-field–unfavored variants of ferroelectric domains (i.e., the [11$\bar{1}$], [$\bar{1}$1$\bar{1}$], [$\bar{1}\bar{1}$1], and [1$\bar{1}$1] domains) is difficult because of the clamping effect from the sample surface (*26*), which severely scatters the light. The results of the phase field simulation under clamping conditions (Fig. 2A and fig. S3) show that some domains with polarization perpendicular to the [011] axis are present to release the elastic energies during the removal of the poling electric field. We also experimentally verified the influence of the clamping effect on the final domain state (movie S1 and fig. S4). We used a polarized light microscope (PLM) to obtain images of a [011]-oriented 0.21PIN-0.47PMN-0.32PT (PIN-PMN-32PT) crystal poled with a conventional poling method, i.e., at room temperature under an electric field of twice

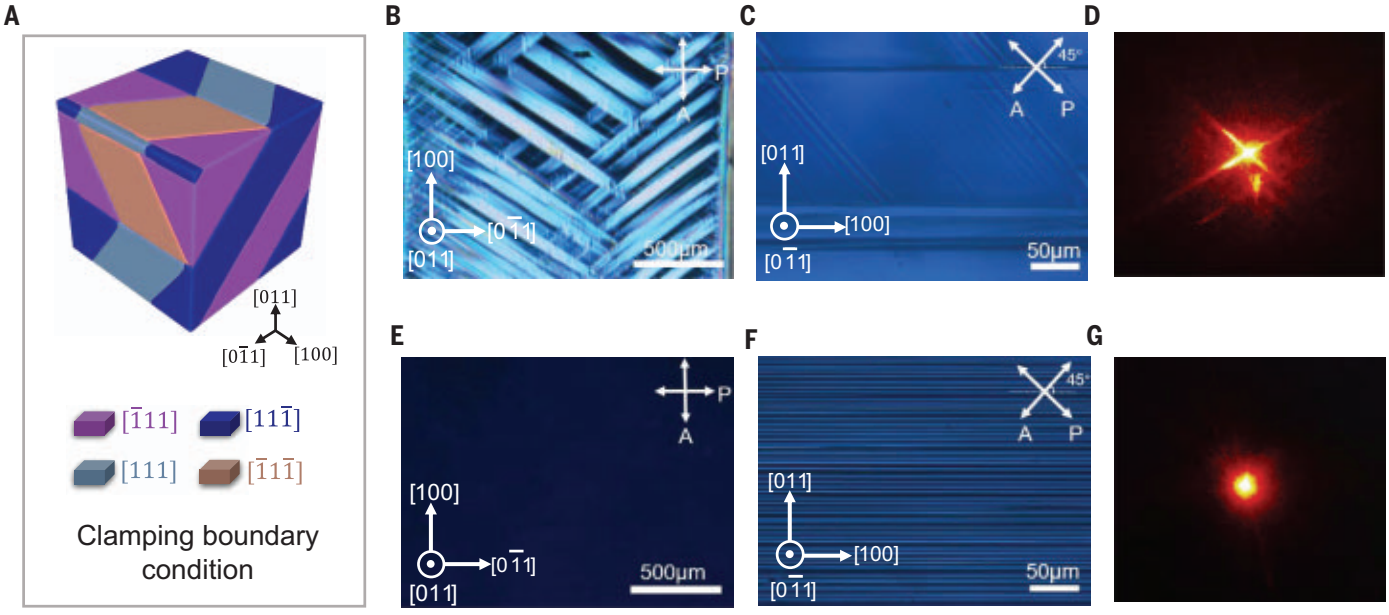


Fig. 2. Domain structures for [011]-oriented rhombohedral PIN-PMN-32PT crystals poled by conventional and high-temperature methods. (**A**) Domain structure of the [011]-poled sample under clamped conditions from phase field simulations. It should be noted that only two domains with polarization perpendicular to the [011] direction are present in Fig. 2A, which is due to the fact that the scale of phase field simulation (64 nm by 64 nm by 64 nm) is much smaller than that of a real sample. We also conducted a phase field simulation involving the other two domains (polarizations are along the $[\bar{1}\bar{1}1]$ and $[\bar{1}1\bar{1}]$ directions) (fig. S3). (**B** and **C**) PLM images on (011) and $(0\bar{1}1)$ surfaces of the [011]-oriented crystals poled by conventional poling method, respectively. (**D**) Output spot of a Gaussian beam which propagates along the [011] direction of the crystal poled by the conventional poling method. (**E** and **F**) PLM images on the (011) and $(0\bar{1}1)$ surfaces, respectively, of the [011]-oriented crystals poled by high-temperature poling method, respectively. The width of the laminar domains of the sample in Fig. 2F is 2.77 μm (SD ~±1.72 μm). (**G**) Output spot of a Gaussian beam that propagates along the [011] direction of the crystal poled by the high-temperature poling method.

the coercive field. (Fig. 2, B and C, and fig. S5). The PLM images taken from the (011) and $(0\bar{1}1)$ surfaces are consistent with the simulated domain structure (Fig. 2A). A number of domain walls are visible in Fig. 2B, in which the angle between the domain walls and $[0\bar{1}1]$ direction is ~35°. The domain walls presented in Fig. 2B are mainly formed by [111] or $[\bar{1}11]$ domain with one of the $[11\bar{1}]/[\bar{1}1\bar{1}]/[\bar{1}11]/[1\bar{1}1]$ ferroelectric domains whose polarization lies in the (011) plane (fig. S6). Because of the presence of these light-scattering domain walls, a 633-nm Gaussian beam with a circular cross section is strongly scattered as the beam goes through the crystal (Fig. 2D). The shape of the scattered beam spot nearly replicates the domain pattern (Fig. 2B) as a result of the diffraction from the refractive indices discontinuity at domain walls (*27*). Evidently, the light scattering will reduce the transparency and adversely affect the efficiency of EO devices.

We employed a high-temperature poling approach to reduce the clamping effect and eliminate the light-scattering domain walls. Specifically, the crystal was heated to ~105° to 115°C and was thus in the orthorhombic phase. Then, by applying an electric field of 5 kV cm^{-1}, we can achieve orthorhombic single-domain configuration. Finally, the crystal was slowly cooled to room temperature with the application of the electric field. During cooling, the orthorhombic single domain with polarization along the [011] direction only split into the rhombohedral [111] and $[\bar{1}11]$ domains. Throughout the process, other variants of rhombohedral domains were absent (movie S2 and fig. S7). By using the above poling strategy, we obtained [011]-poled PIN-PMN-PT crystals without the undesired domains (i.e., $[11\bar{1}]$, $[\bar{1}1\bar{1}]$, $[\bar{1}11]$, or $[1\bar{1}1]$; Fig. 2, E and F, and fig. S8). We observed the extinction phenomenon as the incident light is along the [011] direction with polarization along the $[0\bar{1}1]$ direction (Fig. 2E), whereas the lamellar domain configuration is observed from the $(0\bar{1}1)$ (Fig. 2F) and (100) (fig. S8) planes, which echoes the simulated domain structure under stress-free conditions (Fig. 1A). As expected, the footprint of the Gaussian beam remains as a circular spot after transmitting this crystal along the [011] axis (Fig. 2G), implying substantially enhanced light transmittance compared with that the crystal poled at room temperature. The PIN-PMN-PT crystals used for the discussion of EO properties as well as device design are poled by the high-temperature poling technique.

Transparency and electro-optic properties of the [011]-poled PIN-PMN-PT crystal

We photographed [011]-poled PIN-PMN-PT crystals (Fig. 3A). The poled crystals are highly transparent when viewed from both the [100] and [011] directions. The light transmittance along both directions of the poled sample is ~70% in the wavelength range of 550 to 2500 nm, which is very close to the theoretical limit if only surface reflection is considered (Fig. 3B). Additionally, the [011]-poled PIN-PMN-PT crystal exhibits high transparency in the intermediate infrared band, i.e., 2500 to 5500 nm (Fig. 3C), except for a small absorption peak around the wavelength of 2840 nm resulting from the stretching vibrations of OH^- ions, which was observed in many perovskites such as $SrTiO_3$, $BaTiO_3$, $LiNbO_3$, and $KTaO_3$ (*28*–*30*). We found the optical absorption to monotonically decrease with increasing wavelength (Fig. 3D). As a result of the improved poling technique, the optical transmittance of the present PIN-PMN-PT crystal is considerably higher than that of previously reported relaxor ferroelectric crystals (*14*, *31*–*33*). The transmittance can be further improved through the coating of antireflective film on the transparent surface, e.g., the (100) surface (blue curve in Fig. 3B). In particular, at a wavelength of 1064 nm the optical transmittance is 99.6%, in which the optical absorption of the crystal and surface reflection loss are 0.16 and 0.24%, respectively.

We measured the Pockels coefficients r_{13} and r_{33} with the Mach-Zehnder interferometer (fig. S9A), whereas the effective EO coefficient

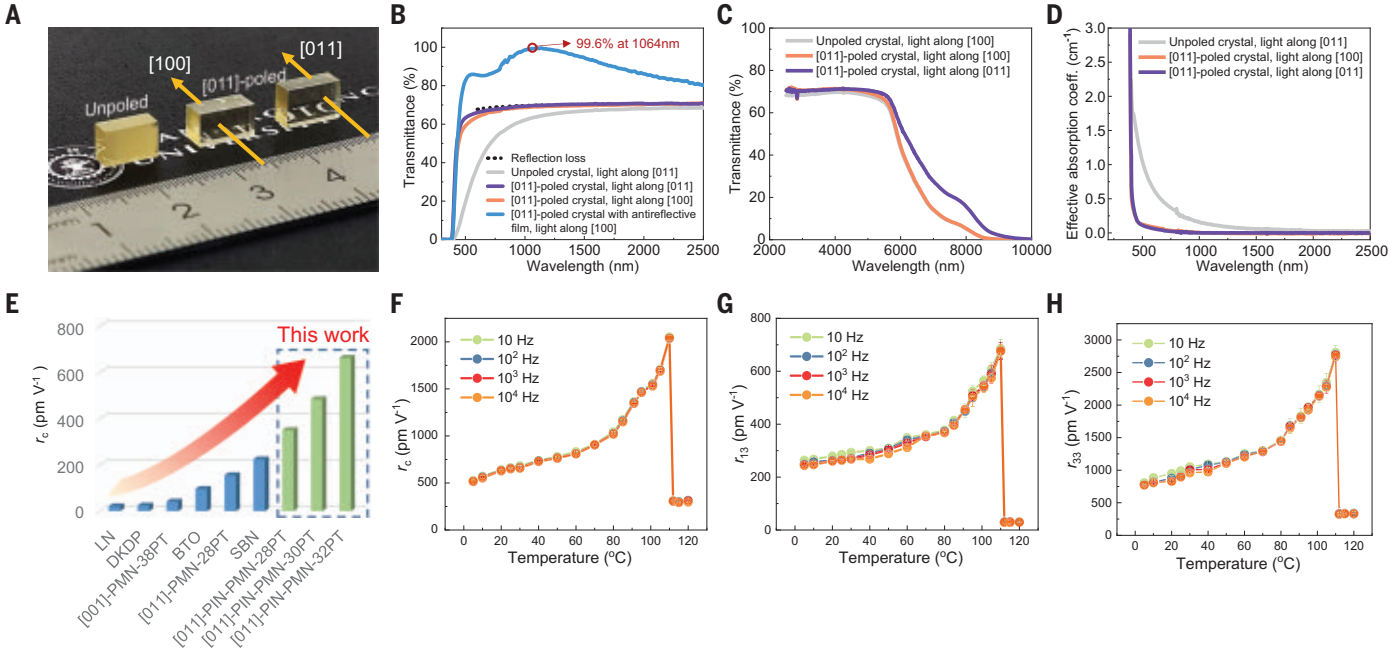


Fig. 3. Transparency and electro-optic properties of PIN-PMN-PT crystals poled along the [011] direction by the high-temperature poling method. Photograph of poled PIN-PMN-PT single crystals with the major faces of (100) and (011), respectively. An unpoled crystal is also given for comparison. The sample size in (**A**) is 8 mm by 3.5 mm by 3 mm. (**B**) Optical transmittance spectra of PIN-PMN-PT samples before and after poling and with antireflective coating (sample thickness is 1.5 mm for optical transmittance measurements). (**C**) Optical transmission spectra of [011]-poled PIN-PMN-PT samples within the wavelength of 2500 to 10,000 nm. (**D**) Effective absorption coefficient of the [011]-poled and unpoled PIN-PMN-PT crystals. (**E**) Comparison of EO coefficient r_c (at room temperature) between [011]-poled PIN-PMN-PT crystals and state-of-the-art EO crystals. (**F** to **H**) Frequency and temperature dependences of r_c, r_{13}, and r_{33}, respectively, for the [011]-poled PIN-PMN-32PT crystal. The error bars indicated the standard errors measured from five points of a sample. It should be noted that EO coefficients can be greatly enhanced as the test frequency approaching the piezoelectric resonance frequency of the sample (fig. S11). This phenomenon is related to the greatly enlarged piezoelectric displacement at the resonance frequency.

r_c was measured with the Senarmont compensator method (fig. S9B) (*34*). We chose a 633-nm He-Ne laser as the light source for the measurement of EO properties. For the rhombohedral PIN-PMN-PT crystals, we found that the EO coefficient r_c increases with increasing PT content, i.e., with the composition approaching the morphotropic phase boundary (MPB) (Fig. 3E). We can attribute this feature to the electric field–induced polarization rotation becoming easier as the composition approaches the MPB (*22*). We mainly focused the following discussion on the PIN-PMN-32PT crystal, which possesses the highest EO coefficient r_c among all studied compositions.

At room temperature (25°C), the r_{33}, r_{13}, and r_c coefficients of the PIN-PMN-32PT crystal were found to be 910, 260, and 670 pm V^{-1}, respectively, outperforming the actively studied EO crystals (Fig. 3E and fig. S10). It is worth noting that the contributions from the variation of the optical path length—which is induced by the converse piezoelectric effect under the applied electric field—to coefficients r_{33}, r_{13}, and r_c are 299, 296, and −4 pm V^{-1}, respectively (*34*). The coefficient r_c is ~30 times as large as that of the state-of-the-art LN and DKDP EO crystals, confirming the potential for use in compact EO devices with low driving voltage. Some other ferroelectric materials may also possess high EO coefficients (r_c > 200 pm V^{-1}), such as (Sr,Ba)Nb_2O_6 crystals (*35*), but these high values are a result of the ferroelectric phase transition temperature being close to room temperature, which greatly limits their operating temperature range. In addition, the EO coefficients of PIN-PMN-32PT crystals are much higher than that of previously reported relaxor ferroelectric crystals, including [011]-poled PMN-28PT (*21*). This is due to the following two reasons: First, the composition of PIN-PMN-32PT is closer to the MPB when compared with the reported PMN-28PT. The electric field-induced polarization rotation becomes easier as the composition approaches the MPB, leading to the higher EO property. Second, by the high-temperature poling technique, the undesired domains in PIN-PMN-32PT crystals can be completely removed without any irreversible electric field-induced phase transition. It is generally believed that the crystal with MPB compositions can be partially transformed into the orthorhombic phase by [011] electric field poling at room temperature because of the over-poling effect, leading to the greatly reduced EO property. By contrast, the applied electric field used in high-temperature poling is much lower than that for room-temperature poling, and thus the PIN-PMN-32PT crystal can fully recover to the rhombohedral phase with the desired domain structure during cooling.

To understand the origin of the giant EO coefficients in rhombohedral PIN-PMN-PT crystals, we performed density functional theory (DFT) calculations. A typical relaxor-PT composition, i.e., 0.75PMN-0.25PT (PMN-25PT), was selected for the calculation as it allows a supercell size feasible for DFT calculations (*36*, *37*); further, its electro-optic and piezoelectric properties are comparable to our studied PIN-PMN-32PT crystal. Through the approach proposed in (*38*–*41*), the intrinsic EO coefficient r_{ijk} under stress-free conditions can be expressed as

$$r_{ijk} = r_{ijk}^{\mathrm{el}} + r_{ijk}^{\mathrm{ion}} + r_{ijk}^{\mathrm{piezo}} \quad (1)$$

The first term r_{ijk}^{el} is a pure electronic part, the second term r_{ijk}^{ion} corresponds to the ionic contribution, and the last term r_{ijk}^{piezo} corresponds to the coupling of piezoelectric and elasto-optic effects. The three contributions were calculated at 0 K for the PMN-25PT crystal. According to our DFT calculations (*34*), the EO coefficients calculated at 0 K (<50 pm V^{-1}, see table S1) are much smaller than those

measured at room temperature. At a temperature far below the phase transition point, the electronic contribution is expected to be insensitive to temperature because it is related to the interaction of the applied electric field with the valence electrons, whereas the ions are confined to their equilibrium positions (*41*). For the PIN-PMN-32PT crystal, the rhombohedral-orthorhombic phase transition temperature and Curie temperature are 105° and 180°C, respectively. Thus, the high EO coefficients at room temperature are thought to be related to the ionic and piezoelectric contributions (i.e., the components r_{ijk}^{ion} and r_{ijk}^{piezo}). These are, in a reasonable approximation, proportional to the dielectric and the piezoelectric tensors, respectively (*34*). The calculated dielectric constant and piezoelectric coefficient of PMN-25PT are lower than 70 and 110 pm V^{-1}, respectively, at 0 K (tables S2 and S3), whereas the measured values are up to 3500 and 1200 pm V^{-1}, respectively, at room temperature, showing a substantial positive correlation with temperature. Therefore, it should be inferred that the components r_{ijk}^{ion} and r_{ijk}^{piezo} increase substantially with temperature as they are responsible for the high room temperature EO coefficients for the PIN-PMN-32PT crystals.

To validate this inference, we measured the EO coefficients of the PIN-PMN-32PT crystal as a function of temperature (Fig. 3, F to H). As the temperature increases from 5 to 60°C, we found that the EO coefficients increase by ~30 to 50%, as they are similar to the temperature-induced changes of the dielectric and piezoelectric coefficients (*22*), revealing that dielectric- and piezoelectric-related contributions play a key role in EO coefficients. The EO coefficients are substantially reduced at a temperature above the rhombohedral-orthorhombic phase transition temperature $T_{r\text{-}o}$ (~105°C), because of both the low piezoelectricity in the orthorhombic phase along the [011] direction and the depoling effect during the phase transition.

In contrast to the strong temperature dependence, the EO coefficients of the PIN-PMN-32PT crystal exhibit excellent stability with respect to frequency. The r_{33}, r_{13}, and r_c coefficients exhibit minor variation below 5% within the frequency range of 10 to 10^4 Hz (Fig. 3, F to H). This feature is vital for EO devices that operate over a broad frequency range, e.g., electro-optic modulators used in the field of broadband radar.

Electro-optic Q-switch made of PIN-PMN-PT crystals

To demonstrate the competitive advantages of our PIN-PMN-PT crystals in practical applications, we constructed and characterized electro-optic Q-switches. Because of the large EO properties and relatively high laser damage threshold (~500 MW cm^{-2}) (*42*), PIN-PMN-PT crystals are promising candidates for Q-switching applications that require ultracompact size and greatly reduced driving voltage.

For benchmark purposes, we also explored key properties of commercial DKDP- and LN-based Q-switches when housed in the same laser cavity for the characterization of the PIN-PMN-32PT Q-switch. We show the commercial DKDP- and LN-based Q-switches and a compact low-voltage Q-switch made of the PIN-PMN-32PT crystal with size 5 mm by 5 mm by 1.5 mm (Fig. 4A). The dimensions of the PIN-PMN-32PT Q-switch are ϕ 12 mm by 3.4 mm, which are considerably smaller than those of the DKDP Q-switch (ϕ 15 mm by 18 mm) and LN Q-switch (ϕ 22 mm by 22 mm); this small size is attributable to its ultrahigh effective EO coefficient, r_c (Fig. 3E). Theoretically, reducing the Q-switch optical travel length is accompanied by an increase in operating voltage, according to Eq. 2

$$V_{\pi/2} = \frac{\lambda d}{2n^3 r_c l} \quad (2)$$

in which $V_{\pi/2}$ is the quarter-wave voltage (i.e., the operating voltage), λ is the wavelength, n is the refractive index, r_c is the effective EO coefficient, l is the length of the crystal along the optical travel direction, and d is the distance between two electrodes of the EO crystal. Because of the ultrahigh EO coefficients, the operating voltage of the studied PIN-PMN-32PT crystal is only 0.2 kV, suggesting 16- and 6.5-fold reductions when compared with that of DKDP and LN, respectively.

We illustrated the laser cavity setup for experimentally characterizing the performance of the PIN-PMN-32PT Q-switch (Fig. 4B). This cavity accommodates a rear mirror, a polarizing beam splitter, an α-cut Nd:YVO_4 acting as the laser host crystal, the Q-switch, a quarter-wave plate, an eighth-wave plate, and an output mirror. The total length of the laser cavity is 65 mm. To ensure that the Q-switch is in high-loss mode under the hold-off state, the position of the eighth-wave plate was carefully adjusted to compensate for the phase retardation induced by the initial birefringence of the PIN-PMN-32PT crystal, leading to zero laser output power. A pulsed laser with a wavelength of 1064 nm is generated at the output end of the cavity through application of a control signal on the PIN-PMN-32PT Q-switch, i.e., pulsed $V_{\pi/2}$ voltage with a repetition rate of 10 Hz to 2 kHz.

We show a single output pulse generated by the PIN-PMN-32PT Q-switch and compared it with commercial DKDP and LN Q-switches at a repetition rate of 1 kHz and pump energy of 3.7 mJ (Fig. 4C). The pulse width of the PIN-PMN-32PT Q-switch is on the order of 1.8 ns—a slight improvement compared with those of commercial Q-switches. The PIN-PMN-32PT Q-switch can produce a more symmetric pulse (Fig. 4C) with a consistently narrower pulse width over a relatively wide repetition rate span (10 Hz to 2 kHz) and pump energy (~2.3 to 3.7 mJ) compared with those of their DKDP- and LN-based counterparts (Fig. 4, D and E). In general, the width of a Q-switched pulse is positively correlated with the intracavity propagation time of the laser beam (*43*). The compact size of the studied PIN-PMN-32PT crystal excessively minimizes the additional optical path length induced by the insertion of the Q-switch (i.e., a product of the geometric length of the crystal along the laser travel direction and the refractive indices of the EO crystal). Consequently, the intracavity propagation time of the PIN-PMN-32PT Q-switched laser is shorter and more strongly favors the narrowing of the output pulse width.

The PIN-PMN-32PT Q-switched laser exhibits a satisfactory output pulse energy and optical-to-optical efficiency (~7.7%) at a repetition rate of 1 kHz (Fig. 4F), both of which are highly comparable to those of the laser pulses generated by the commercial DKDP- and LN-based Q-switches, with efficiency on the order of 7.9 and 8.0%, respectively.

Peak power, represented by the pulse energy and pulse width, is another key parameter for understanding the performance of a Q-switch. The maximum peak power of the pulse produced by the studied PIN-PMN-32PT Q-switch was calculated to be 154 kW at input pumping energy of 3.7 mJ, which is nearly identical to that of the DKDP- and LN-based Q-switched laser (Fig. 4G and table S4). This substantial peak power is primarily ascribed to the shorter pulse width, which effectively offsets the minor deficiency in the output pulse energy. This result provides further evidence that the studied PIN-PMN-32PT crystals are of extremely high quality in terms of optical transparency and homogeneity and thus meet the standards of a commercial product.

The pulse-to-pulse energy stability of the pulse train is also an important feature for evaluating Q-switch performance. We recorded the energy profile of 100 sequential output pulses produced by the PIN-PMN-32PT Q-switch at a pump energy of 3.7 mJ and a repetition rate of 1 kHz (Fig. 4H). The variation coefficient of the laser output energy over the pulse train is estimated to be 2.7%, demonstrating an ultralow output energy jitter on par with that of the commercial DKDP and LN Q-switches (Fig. 4I and fig. S12).

Through ferroelectric phase, crystal orientation, and poling process designs, we successfully boosted the transparency of PIN-PMN-PT crystals (99.6% at a wavelength of 1064 nm) and achieved very high EO coefficients (r_{33} and r_c of 900 and 670 pm V^{-1}, respectively). Because the PIN-PMN-PT crystals have desirable electro-optic properties, we used them for

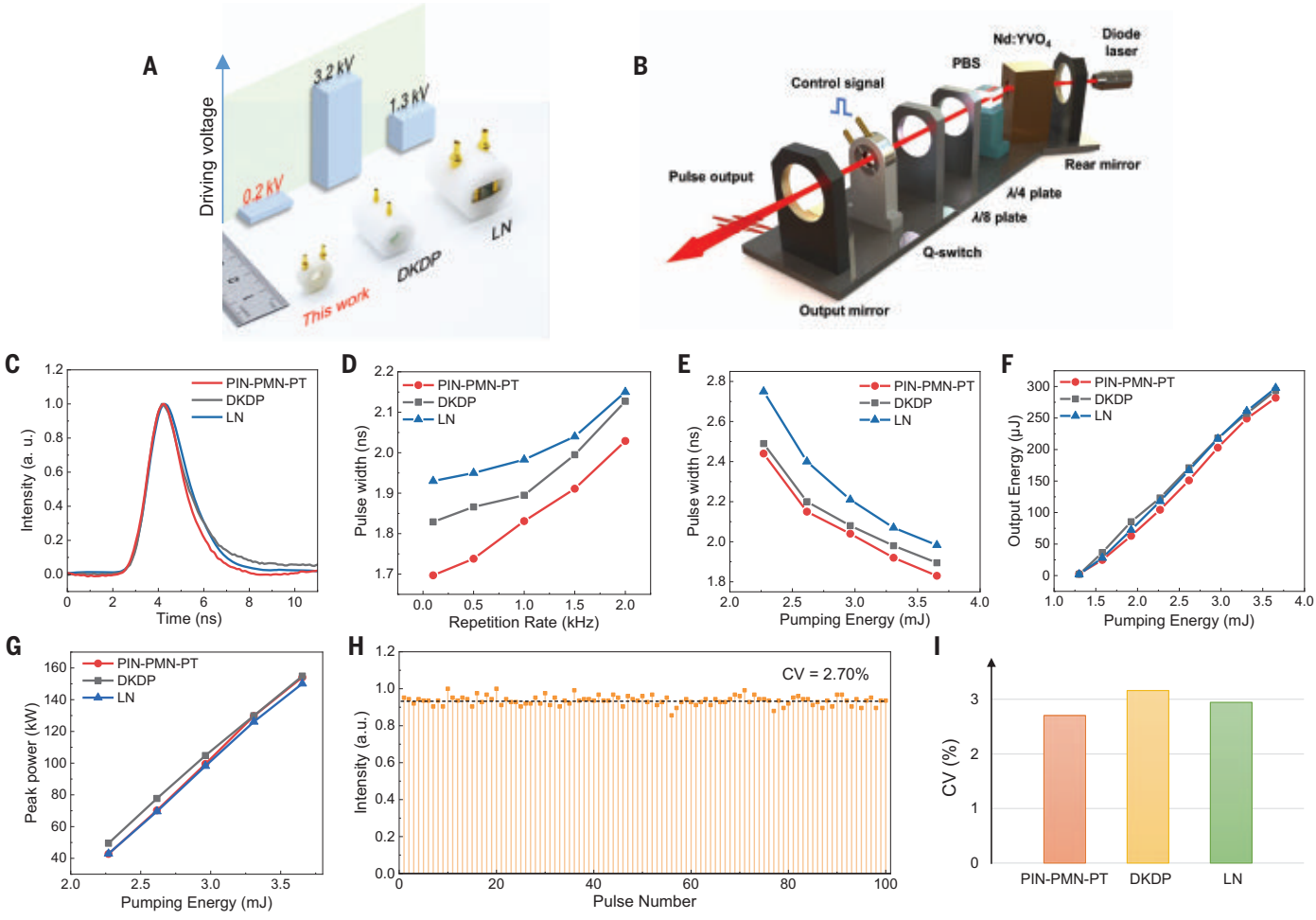


Fig. 4. Key performances of the Q-switch fabricated by the transparent PIN-PMN-32PT crystal. (**A**) EO Q-switch based on different crystals, in which operating voltages $V_{\pi/2}$ (at 1064 nm) are given as well. For PIN-PMN-32PT and LN-based Q-switches, the light travel direction and applied electric field direction are perpendicular, whereas for the DKDP-based Q-switch, the light and electric field are along the same direction. (**B**) Schematic of Nd:YVO_4 laser Q-switched by PIN-PMN-32PT in the pulse-on cavity at λ = 1064 nm. (**C**) Single pulse profile of the oscillator at 1 kHz and 3.7 mJ. (**D**) Pulse width as function of repetition rate (pumping energy of 3.7 mJ). (**E**) Pulse width as a function of pumping energy (repetition rate of 1 kHz). (**F**) Output pulse energy at 1 kHz as a function of pumping energy. (**G**) Peak power as a function of pumping energy. (**H**) Recorded output pulse train consisting of 100 sequential pulses at 1 kHz, showing an ultralow energy jitter of PIN-PMN-PT-based Q-switch (CV, ratio of standard deviation to the mean value). (**I**) The CV of PIN-PMN-PT, DKDP and LN crystals, measured at 1 kHz.

the design of a free-space Q-switch. Compared with the commercial LN and DKDP-based Q-switches, the volume of the PIN-PMN-32PT-based Q-switch is substantially reduced by more than one order of magnitude, with the operating voltage reduced to 200 V from a starting point of 1300 to 3200 V. Given the fact that the Q-switch is a key component of pulsed lasers, the miniaturization and low driving voltage of the Q-switches will considerably reduce the size, weight, and power consumption of pulsed lasers while also mitigating issues relating to electromagnetic interference induced by high-voltage pulses. This development will benefit numerous applications, such as ultracompact and low-power–consumption laser radars, enable sensing functions in navigation with small-scale robots and intelligent recognition in autonomous driving, and allow for improved precision in medical and scientific equipment requiring high stability and reliability.

REFERENCES AND NOTES

1. M. Zhang *et al.*, *Nature* **568**, 373–377 (2019).
2. G. T. Reed, G. Mashanovich, F. Y. Gardes, D. J. Thomson, *Nat. Photonics* **4**, 518–526 (2010).
3. J. D. Bull, N. A. Jaeger, H. Kato, M. Fairburn, A. Reid, P. Ghanipour, "40-GHz electro-optic polarization modulator for fiber optic communications systems" in *Proc. SPIE 5577, Photonics North 2004: Optical Components and Devices* (Society of Photo-Optical Instrumentation Engineers, 2004), pp. 133–143.
4. C. Wang *et al.*, *Nature* **562**, 101–104 (2018).
5. L. R. Dalton, P. A. Sullivan, D. H. Bale, *Chem. Rev.* **110**, 25–55 (2010).
6. E. L. Wooten *et al.*, *IEEE J. Sel. Top. Quantum Electron.* **6**, 69–82 (2000).
7. H. Nakano, K. Kanz, C. A. Ebbers, "A thermally compensated, deuterated KDP Q-switch for high average power lasers" in *Summaries of Papers Presented at the Lasers and Electro-Optics. CLEO '02. Technical Diges. IEEE*, (2002) pp. 179–180.
8. G. Li *et al.*, *J. Cryst. Growth* **274**, 555–562 (2005).
9. B. A. Fuchs, P. P. Hed, P. C. Baker, *Appl. Opt.* **25**, 1733–1735 (1986).
10. M. Aillerie, N. Théofanous, M. D. Fontana, *Appl. Phys. B* **70**, 317–334 (2000).
11. M. Zgonik *et al.*, *Phys. Rev. B Condens. Matter* **50**, 5941–5949 (1994).
12. M. Zgonik, R. Schlesser, I. Biaggio, P. Günter, *Ferroelectrics* **158**, 217–222 (1994).
13. X. Wan *et al.*, *Appl. Phys. Lett.* **85**, 5233–5235 (2004).
14. X. Wan, H. Luo, J. Wang, H. L. W. Chan, C. L. Choy, *Solid State Commun.* **129**, 401–405 (2004).
15. Y. Zhao *et al.*, *Ferroelectrics* **542**, 112–119 (2019).
16. Y. Zhao *et al.*, *J. Appl. Phys.* **123**, 084104 (2018).
17. D.-Y. Jeong, Y. Lu, V. Sharma, Q. Zhang, H.-S. Luo, *Jpn. J. Appl. Phys.* **42**, 4387–4389 (2003).
18. Q. Hu *et al.*, *Appl. Phys. Lett.* **115**, 222901 (2019).
19. C. Qiu *et al.*, *Nature* **577**, 350–354 (2020).
20. F. Li, L. Wang, L. Jin, Z. Xu, S. Zhang, *Cryst. Eng. Comm.* **16**, 2892–2897 (2014).
21. C. Deng *et al.*, *Adv. Mater.* **33**, e2103013 (2021).
22. S. Zhang, F. Li, *J. Appl. Phys.* **111**, 031301 (2012).
23. H. Fu, R. E. Cohen, *Nature* **403**, 281–283 (2000).
24. F. Li *et al.*, *Nat. Commun.* **7**, 13807 (2016).

25. S. Zhang *et al*., *Prog. Mater. Sci.* **68**, 1–66 (2015).
26. Z. Chen *et al*., Giant tuning of ferroelectricity in single crystals by thickness engineering. *Sci. Adv.* **6**, eabc7156 (2020).
27. M. Müller, E. Soergel, M. C. Wengler, K. Buse, *Appl. Phys. B* **78**, 367–370 (2004).
28. F. G. Wakim, *J. Chem. Phys.* **49**, 3738–3739 (1968).
29. A. Jovanović, M. Wöhlecke, S. Kapphan, A. Maillard, G. Godefroy, *J. Phys. Chem. Solids* **50**, 623–627 (1989).
30. L. Kovács, M. Wohlecke, A. Jovanović, K. Polgár, S. Kapphan, *J. Phys. Chem. Solids* **52**, 797–803 (1991).
31. C. He *et al*., *J. Appl. Phys.* **110**, 083513 (2011).
32. F. Wu *et al*., *J. Mater. Sci.* **47**, 2818–2822 (2012).
33. E. Sun *et al*., *J. Appl. Phys.* **107**, 113532 (2010).
34. Materials and methods are available as supplementary materials.
35. S. Ducharme, J. Feinberg, R. Neurgaonkar, *IEEE J. Quantum Electron.* **23**, 2116–2121 (1987).
36. H. Takenaka, I. Grinberg, S. Liu, A. M. Rappe, *Nature* **546**, 391–395 (2017).
37. F. Li *et al*., *Science* **364**, 264–268 (2019).
38. M. Veithen, X. Gonze, Ph. Ghosez, *Phys. Rev. B Condens. Matter Mater. Phys.* **71**, 125107 (2005).
39. A. K. Hamze, A. A. Demkov, *Phys. Rev. Mater.* **2**, 115202 (2018).
40. A. K. Hamze, M. Reynaud, J. Geler-Kremer, A. A. Demkov, *Npj Comput. Mater.* **6**, 130 (2020).
41. M. Veithen, Ph. Ghosez, *Phys. Rev.* **71**, 132101 (2005).
42. W. Zhao *et al*., *Ceram. Inter.* **42**, 11909–11914 (2022)
43. W. Koechner, *Solid-State Laser Engineering*, vol. 1 (Springer, 1988).

ACKNOWLEDGMENTS

Funding: F.L., Z.X., and X.Y.W. acknowledge the support of the National Natural Science Foundation of China (grants 51922083, 51831010, and 51761145024), the development programme of Shaanxi province (grants 2019ZDLGY04-09), the National Key R&D Program of China (grant 2021YFE0115000), and the 111 Project (B14040). H.T. acknowledge the support of the National Natural Science Foundation of China (grants 12074092 and 12004085) and the fellowship of China National Postdoctoral Program for Innovative Talents (grant BX20200111). B.X. acknowledges financial support from the National Natural Science Foundation of China (grant 12074277) and the Natural Science Foundation of Jiangsu Province (BK20201404). D.W. acknowledges the support of the Australian Research Council (FT180100541). B.W. and L.Q.C. acknowledge the support of the US National Science Foundation under grant DMR-1744213 and Materials Research Science and Engineering Center (MRSEC) grant DMR-1420620. **Author contributions:** The work was conceived and designed by S.Z., H.T., and F.L.; X.L. prepared the samples, with assistance from C.Q. and L.Q.; X.L. and P.T. performed the optical experiments, with assistance from W.Z.; P.T. and X.J. fabricated the device and characterized the performances, with assistance from C.W. and X.L.; K.S. and H.G. grew the crystals; F.L. and Z.X. supervised crystal growth; F.L., X.W., and H.T. supervised the optical experiments; H.T. supervised device fabrication and characterization; X.M. and B.X. performed the DFT calculation; Y.L. and B.W. performed the phase field simulations with the supervision of F.L. and L.Q.C.; F.L. and X.L. drafted the manuscript; S.Z., D.W., S.L., and L.Q.C. revised the manuscript, and all authors discussed the results. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All relevant data are available in the main text or the supplementary materials.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abn7711
Materials and Methods
Supplementary Text
Figs. S1 to S12
Tables S1 to S7
References (*44*–*63*)
Movies S1 and S2

19 December 2021; accepted 23 March 2022
10.1126/science.abn7711

CELL BIOLOGY

ESCRT-mediated membrane repair protects tumor-derived cells against T cell attack

Alex T. Ritter[1]*, Gleb Shtengel[2], C. Shan Xu[2], Aubrey Weigel[2], David P. Hoffman[2]†, Melanie Freeman[2]†, Nirmala Iyer[2], Nensi Alivodej[2]‡, David Ackerman[2], Ilia Voskoboinik[3], Joseph Trapani[3], Harald F. Hess[2], Ira Mellman[1]*

Cytotoxic T lymphocytes (CTLs) and natural killer cells kill virus-infected and tumor cells through the polarized release of perforin and granzymes. Perforin is a pore-forming toxin that creates a lesion in the plasma membrane of the target cell through which granzymes enter the cytosol and initiate apoptosis. Endosomal sorting complexes required for transport (ESCRT) proteins are involved in the repair of small membrane wounds. We found that ESCRT proteins were precisely recruited in target cells to sites of CTL engagement immediately after perforin release. Inhibition of ESCRT machinery in cancer-derived cells enhanced their susceptibility to CTL-mediated killing. Thus, repair of perforin pores by ESCRT machinery limits granzyme entry into the cytosol, potentially enabling target cells to resist cytolytic attack.

Cytotoxic lymphocytes, including cytotoxic T lymphocytes (CTLs) and natural killer (NK) cells, are responsible for identifying and destroying virus-infected or tumorigenic cells. To kill their targets, CTLs and NK cells secrete a pore-forming toxin called perforin through which apoptosis-inducing serine proteases (granzymes) are delivered directly into the cytosol. Successful killing of target cells often requires multiple hits from single or multiple T cells (*1*). This has led to the idea that cytotoxicity is additive, often requiring multiple rounds of sublethal lytic granule secretion events before a sufficient threshold of cytosolic granzyme activity is reached to initiate apoptosis in the target (*2*).

Loss of plasma membrane integrity induced by cytolytic proteins or mechanical damage leads to a membrane repair response. Damage results in an influx of extracellular Ca^{2+}, which has been proposed to lead to the removal of the membrane lesion by endocytosis, resealing of the lesions by lysosomal secretion, or budding into extracellular vesicles (*3*). Perforin pore formation was initially reported to enhance endocytosis of perforin (*4*), but subsequent work has challenged this claim (*5*). Endosomal sorting complexes required for transport (ESCRT) proteins can repair small wounds and pores in the plasma membrane caused by bacterial pore-forming toxins, mechanical wounding, and laser ablation (*6*, *7*). ESCRT proteins are transiently recruited to sites of membrane damage in a Ca^{2+}-dependent fashion, where they assemble budding structures that shed to eliminate the wound and restore plasma membrane integrity. ESCRT-dependent membrane repair has been implicated in the resealing of endogenous pore-mediated plasma membrane damage during necroptosis (*8*) and pyroptosis (*9*).

[1]Genentech, Inc., South San Francisco, CA 94080, USA. [2]Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA 20147, USA. [3]Rosie Lew Cancer Immunology Program, Peter MacCallum Cancer Centre, Melbourne VIC, Australia.
***Corresponding author. Email: ritter.alex@gene.com (A.T.R.); iram@gene.com (I.M.)**
†Present address: 10X Genomics, Inc., Pleasanton, CA 94566, USA.
‡Present address: Neurovascular Interface, Goethe University Frankfurt and Max Planck Institute for Brain Research, 60438 Frankfurt am Main, Germany.

Localization of target-derived ESCRT proteins to the cytolytic synapse

To investigate whether ESCRT-mediated membrane repair might be involved in the removal of perforin pores during T cell killing, we first determined whether ESCRT proteins in cancer-derived cells were recruited to sites of CTL engagement after perforin secretion. We used CTLs from OT-I mice that express a high-affinity T cell receptor (TCR) that recognizes the ovalbumin peptide SIINFEKL ($OVA_{257-264}$) bound to the major histocompatibility complex (MHC) allele H-2Kb (*10*). We performed live-cell microscopy of OT-I CTLs engaging SIINFEKL-pulsed target cells that express enhanced green fluorescent protein (EGFP)–tagged versions of Tsg101 or Chmp4b, two ESCRT proteins implicated in membrane repair (*6*). To correlate recruitment of ESCRT proteins with perforin exposure in time, we monitored CTL-target interaction in media with a high concentration of propidium iodide (PI), a cell-impermeable fluorogenic dye that can rapidly diffuse through perforin pores to bind and illuminate nucleic acids in the cytosol and nucleus of the target (*5*). EGFP-tagged ESCRT proteins were consistently recruited to the site of CTL engagement within 30 to 60 s after PI influx (Fig. 1, A and B). EGFP-Tsg101 and EGFP-Chmp4b in target cells accumulated at the cytolytic synapse after PI influx in 25 of 27 (92.6%) and 31 of 33 (93.9%) of conjugates monitored, respectively, compared with a cytosolic EGFP control, which was not recruited

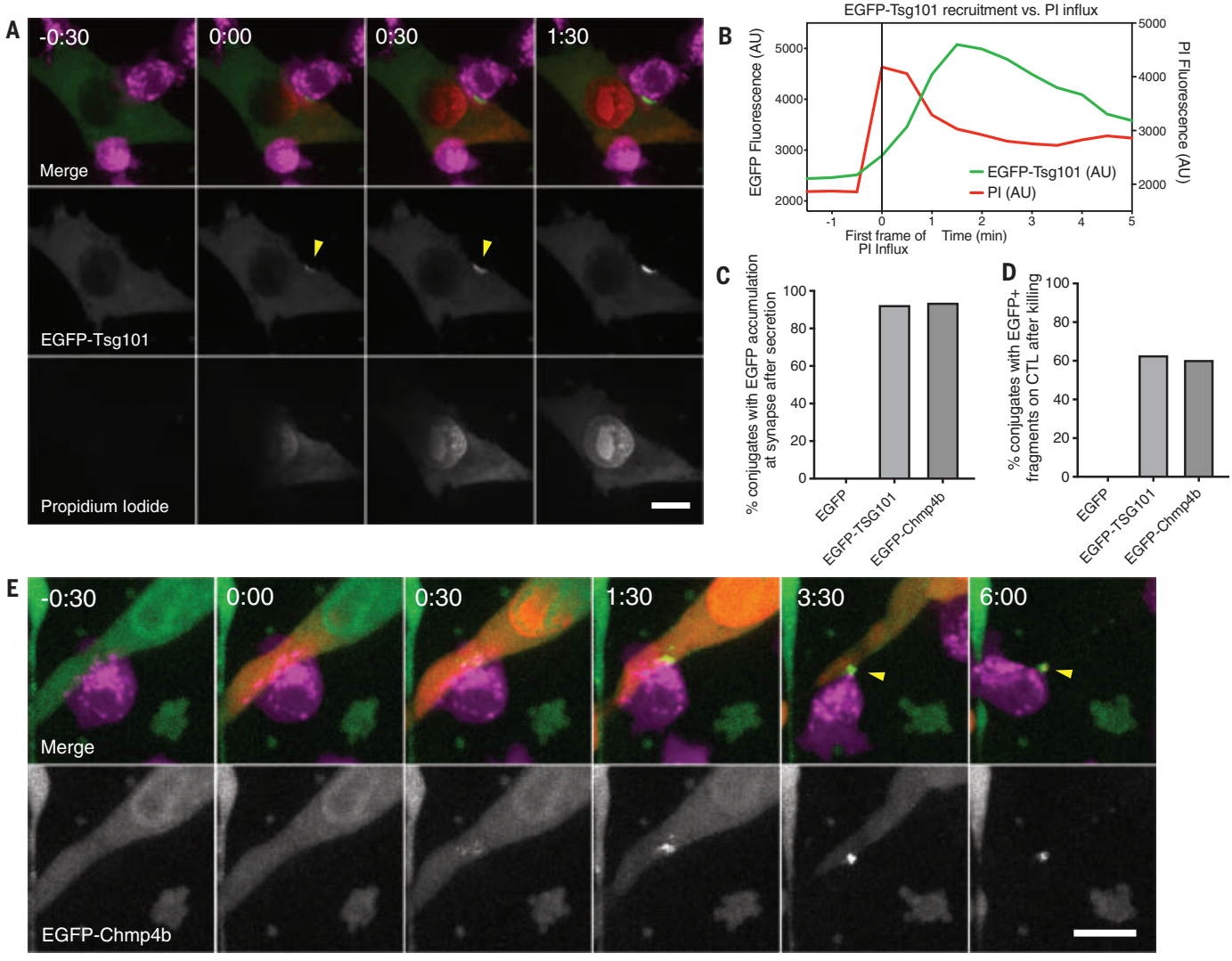


Fig. 1. Fluorescently tagged ESCRT proteins in targets localize to site of CTL killing after perforin secretion. (**A**) Live-cell spinning disk confocal imaging of a fluorescently labeled OT-I CTL (magenta) engaging an MC38 cancer cell expressing EGFP-Tsg101 (green) in media containing 100 μM PI (red). Yellow arrowheads highlight ESCRT recruitment. T-0:00 is the first frame of PI influx into the target cell (time in minutes:seconds). Scale bar, 10 μm. (**B**) Graph of EGFP-Tsg101 and PI fluorescence intensity at the IS within the target over time, from example in (A). AU, arbitrary units. (**C** and **D**) Quantification of CTL-target conjugates exhibiting accumulation of EGFP at the synapse after PI influx (C) or detectable EGFP-labeled material associated with CTL after target interaction (D) (EGFP condition: *N* = 22 conjugates in seven independent experiments; EGFP-Tsg101 condition: *N* = 27 conjugates in nine independent experiments; EGFP-Chmp4b condition: *N* = 33 conjugates in 24 independent experiments). (**E**) Live-cell spinning disk confocal imaging of OT-I CTL (magenta) killing MC38 expressing EGFP-Chmp4b (green), demonstrating the presence of target-derived EGFP-Chmp4b material (yellow arrowheads) associated with CTL membrane after a productive target encounter. T-0:00 is the first frame of PI influx into the target cell. Scale bar, 10 μm.

(Fig. 1C and movies S1 to S3). Notably, ESCRT-laden material, presumably membrane fragments, frequently detached from the target cell and adhered to the surface of the CTL (Fig. 1, D and E, and movie S2). We observed this phenomenon in ~60% of conjugates imaged in which targets expressed EGFP-Tsg101 or EGFP-Chmp4b (17 of 27 and 20 of 33 conjugates, respectively; Fig. 1D). Shedding of ESCRT-positive membrane from the cell after repair occurs after laser-induced plasma membrane wounding (*6*, *7*). Plasma membrane fragments shed from the target cell into the synaptic cleft likely contain ligands for CTL-resident receptors. Target cell death would separate the CTL and target, revealing target-derived material on the CTL surface.

3D cryo-SIM and FIB-SEM imaging of CTLs caught in the act of killing target cells

Although live-cell imaging indicated that ESCRT complexes were rapidly recruited at sites of T cell–target cell contact, light microscopy alone is of insufficient resolution to establish that this event occurred at the immunological synapse (IS). We thus sought to capture a comprehensive view of the IS in the moments immediately after secretion of lytic granules. We used cryo–fluorescence imaging followed by correlative focused ion beam–scanning electron microscopy (FIB-SEM), which can achieve isotropic three-dimensional (3D) imaging of whole cells at 8-nm resolution or better (*11*–*13*). To capture the immediate response of target cells after perforin exposure, we developed a strategy whereby cryo-fixed CTL-target conjugates were selected shortly after perforation, indicated by the presence of a PI gradient in the target (fig. S1A). In live-cell imaging

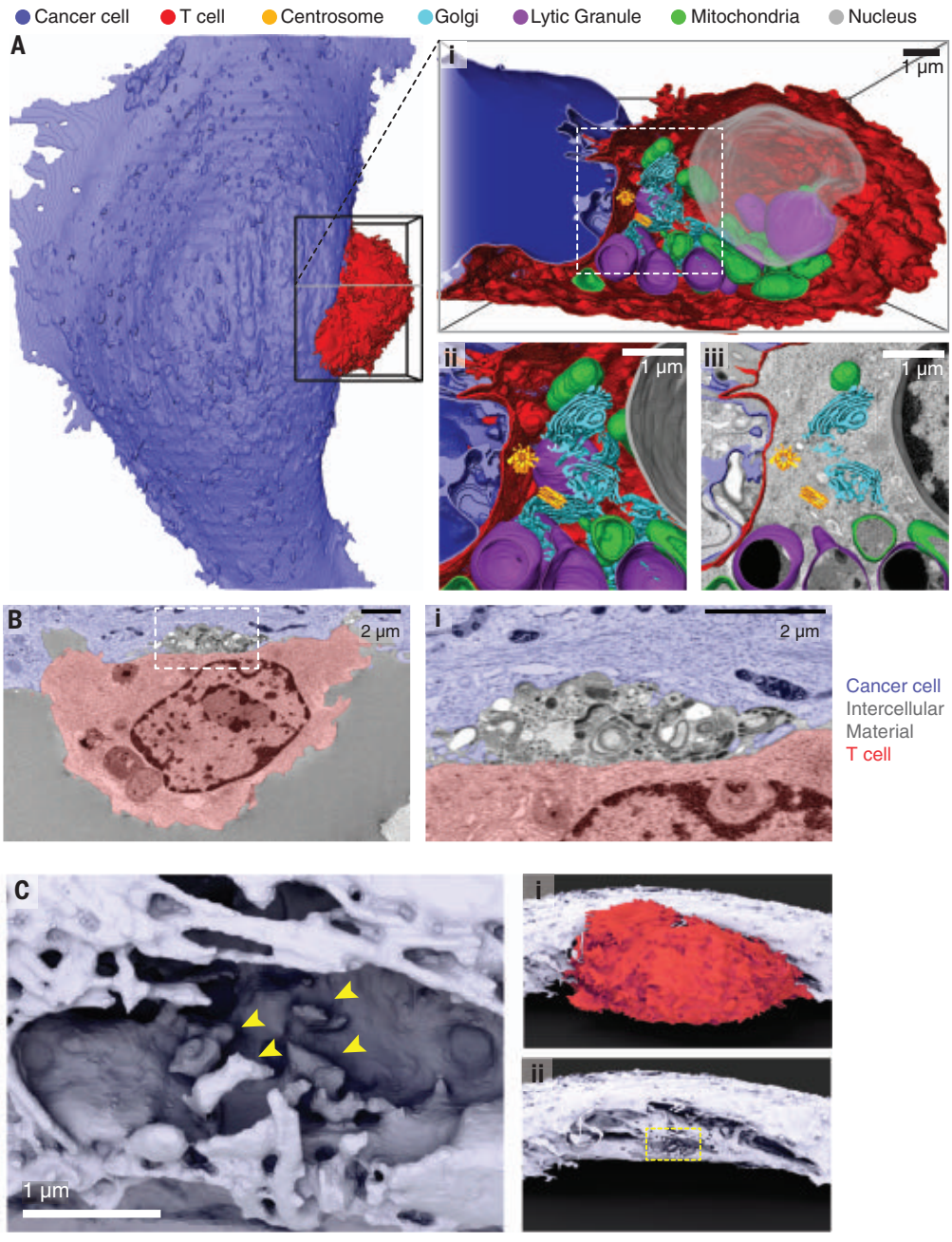


Fig. 2. Eight-nm-resolution 3D FIB-SEM imaging of whole CTL-target conjugate. (**A**) 3D rendering of segmented plasma membrane predictions derived from isotropic 8-nm-resolution FIB-SEM imaging of a high-pressure frozen OT-I CTL (red) captured moments after secretion of lytic granules toward a peptide-pulsed ID8 ovarian cancer cell (blue). (i) Side-on sliced view corresponding to the gray horizontal line within the inset box in (A). Seen here are 3D renderings of the segmented plasma membrane of the cancer cell (blue) as well as the CTL plasma membrane (red), centrosome (gold), Golgi apparatus (cyan), lytic granules (purple), mitochondria (green), and nucleus (gray). (ii and iii) A zoomed-in view from the dashed white box in (i) shows the details of the IS (ii) and a single corresponding FIB-SEM slice docked onto the segmented data (iii). (**B**) Single top-down FIB-SEM slice showing overlaid target cell (blue) and CTL (red) segmentation. (i) Zoomed-in view from dashed white box in (B) details the intercellular material (IM) (gray) between the CTL and target at the IS. (**C**) Zoomed-in image of a 3D rendering of the surface of the target cell plasma membrane (white) opposite the intercellular material (IM) at the IS. Yellow arrowheads mark plasma membrane buds protruding into the synaptic cleft. (i and ii) Accompanying images demonstrate the orientation of the view in (C) with the rendering of the CTL (red) present (i) and removed (ii), and the dashed yellow box in (ii) indicates the area of detail shown in (C).

experiments, PI fluorescence across the nucleus of SIINFEKL-pulsed ID8 target cells began as a gradient and became homogeneous 158 ± 64 s, on average, after initial PI influx (N = 31 conjugates; fig. S1, B and C, and movie S4). Thus, fixed CTL-target conjugates that exhibited a gradient of PI across the nucleus would have been captured within ~3 min of perforin exposure.

Coverslips of CTL-target conjugates underwent high-pressure freezing and were subsequently imaged with wide-field cryogenic fluorescence microscopy followed by 3D cryo–structured illumination microscopy (3D cryo-SIM) performed in a customized optical cryostat (*14*). We selected candidate conjugates for FIB-SEM imaging on the basis of whether a gradient of PI fluorescence was observed across the nucleus of the target emanating from an attached CTL (movie S5). FIB-SEM imaging of the CTL-target conjugate at 8-nm isotropic voxels resulted in a stack of >10,000 individual electron microscopy (EM) images. The image stack was then annotated using a human-assisted machine learning–computer vision platform to segment the plasma membranes of each cell along with cell nuclei and various organelles (https://ariadne.ai/).

We captured four isotropic 3D 8-nm-resolution EM datasets of CTLs killing cancer cells moments after the secretion of lytic granule contents (Fig. 2A and movie S6). Semiautomated segmentation of the cell membranes, nuclei, lytic granules, Golgi apparatus, mitochondria, and centrosomes of the T cells allow for easier visualization and analysis of the 3D EM data. All FIB-SEM datasets and segmentations can be explored online at https://openorganelle.janelia.org (see links in

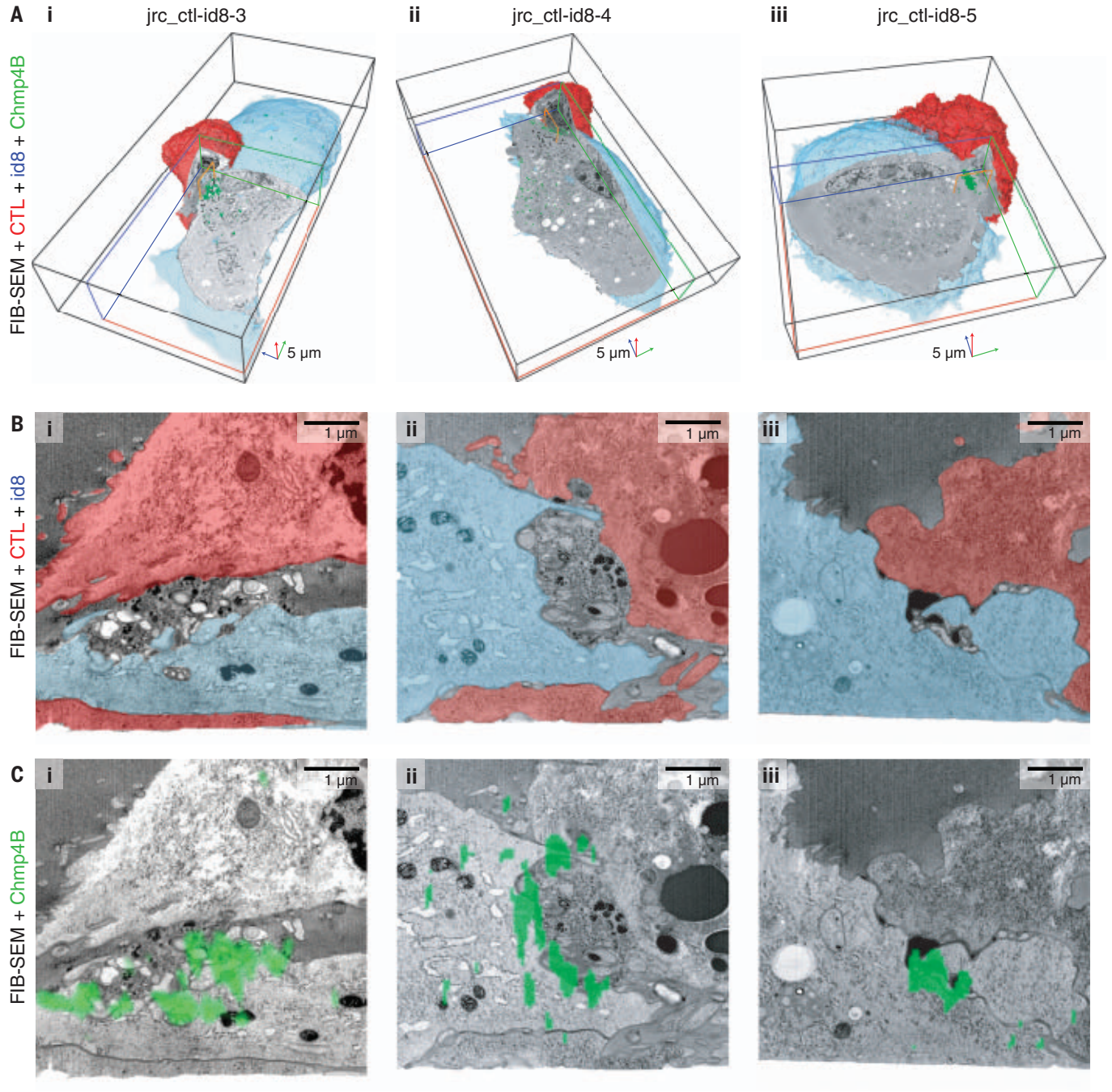


Fig. 3. Correlative 3D cryo-SIM and FIB-SEM reveal localization of target-derived ESCRT within the cytolytic IS. (**A**) Three example datasets showing correlative 3D cryo-SIM and FIB-SEM imaging of OT-I CTLs (red) captured moments after secretion of lytic granules toward peptide-pulsed ID8 cancer cells (blue) expressing mEmerald-Chmp4b (green fluorescence). (**B** and **C**) Single FIB-SEM slices corresponding to the orange boxes in (A), overlaid with CTL and cancer cell segmentation (B) or correlative cryo-SIM fluorescence of mEmerald-Chmp4b derived from the target cell (C).

the supplementary materials). Reconstructed views of the segmented data clearly demonstrate the polarization of the centrosome, Golgi apparatus, and lytic granules to the IS—all of which are hallmarks of CTL killing [Fig. 2A, i to iii, and movie S6, time stamp (TS) 1:33] (*15*, *16*). On the target cell side, we noted cytoplasmic alterations consistent with cell damage including enhanced electron density of mitochondria adjacent to the IS (fig. S2A). Close visual scanning of the postsynaptic target cell membrane in the raw EM data failed to reveal obvious perforin pores, which have diameters (16 to 22 nm) close to the limit of resolution for this technique (*17*).

The segmentation of the two cells illustrates the detailed topography of the plasma membrane of the CTL and target at the IS (fig. S2B). The raw EM and segmentation data reveal a dense accumulation of particles, vesicles, and multilamellar membranous materials, which crowd the synaptic cleft between the CTL and the target (Fig. 2B and movie S6, TS 0:40 to 0:50). The source of this intercellular material (IM) was likely in part the lytic granules because close inspection revealed similar particles and dense vesicles located

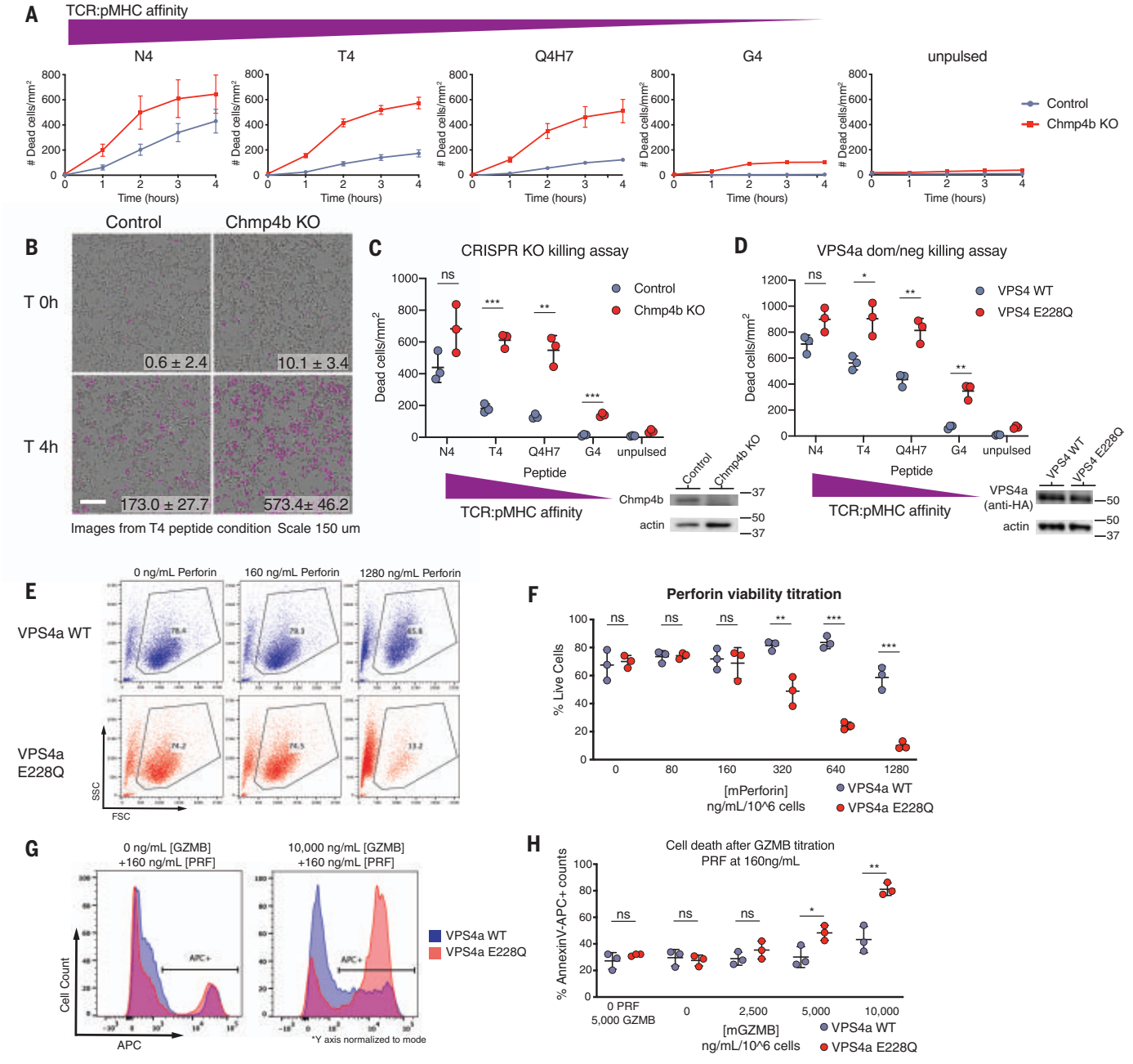


Fig. 4. ESCRT inhibition enhances susceptibility of cancer cells to CTL killing and recombinant lytic proteins. (**A**) Representative time-lapse data of killing of peptide-pulsed Chmp4b knockout (KO) or control B16-F10 cells by OT-I CTLs. Affinity of the pulsed peptide to OT-I TCR decreases from left to right. Error bars indicate SDs. (**B**) Images extracted from T4 medium-affinity peptide condition show software-detected caspase 3/7+ events in control and Chmp4b KO conditions. (**C** and **D**) Data representing the 4-hour time point of assays measuring OT-I T cells killing either Chmp4b KO (C) or VPS4 dominant-negative (D) target cells with matched controls. Error bars indicate SDs of data. Data are representative of at least three independent experimental replicates. pMHC, peptide-MHC; HA, hemagglutinin. (**E** and **F**) Determination of sublytic dose of Prf. B16-F10 cells expressing VPS4a (WT or E228Q) were exposed to increasing concentrations of Prf. Cell viability was determined by morphological gating (E). FSC, forward scatter; SSC, side scatter. (**G** and **H**) B16-F10 cells expressing VPS4a (WT or E228Q) were exposed to a sublytic dose of Prf in combination with increasing concentrations of recombinant GZMB (rGZMB). Cell death was determined by Annexin V–allophycocyanin (APC) staining (G). Controls include a condition with no perforin and 5000 ng/ml rGZMB and sublytic perforin with no rGZMB. Graphs in (F) and (H) represent the means of three experiments, and error bars indicate SDs. Statistical significance was determined by multiple unpaired *t* tests with alpha = 0.05. ns, not significant; $*P < 0.05$; $**P < 0.01$; $***P < 0.001$.

within as-yet-unreleased granules (fig. S2C). To determine whether some of the membranous material within the intercellular space might also have been derived from the target cell, we examined the surface topology of the postsynaptic target cell. We noted multiple tubular and bud-like protrusions of the target cell membrane that extended into the synaptic space; thus, at least some of the membrane structures observed were still in continuity with the target cell (Fig. 2C and movie S6, TS 0:58 to 1:11). ESCRT proteins have been shown

to generate budding structures in the context of plasma membrane repair (*6*), which led us to next assess where target-derived ESCRT proteins are distributed in the context of the postsecretion IS.

To map the localization of target-derived ESCRT proteins onto a high-resolution landscape of the IS, we captured three FIB-SEM datasets that have associated 3D cryo-SIM fluorescence data for mEmerald-Chmp4b localization (Fig. 3A, fig. S3, and movie S7). This correlative light and electron microscopy (CLEM) revealed that mEmerald-Chmp4b expressed in the target cell was specifically recruited to the target plasma membrane opposite the secreted IM (Fig. 3, B and C). The topography of the plasma membrane at the site of ESCRT recruitment was markedly convoluted, exhibiting many bud-like projections (movie S7, TS 0:37 to 0:40). mEmerald-Chmp4b fluorescence also overlapped with some vesicular structures in the intercellular synaptic space (Fig. 3C). Together, the live-cell imaging and the 3D cryo-SIM and FIB-SEM CLEM demonstrate the localization of ESCRT proteins at the synapse that was the definitive site of CTL killing and was thus spatially and temporally correlated to perforin secretion. These data implicate the ESCRT complex in the repair of perforin pores.

Function of ESCRT proteins in repair of perforin pores

We next investigated whether ESCRT inhibition could enhance the susceptibility of target cells to CTL-mediated killing. Prolonged inactivation of the ESCRT pathway is itself cytotoxic (*9*). We thus developed strategies to ablate ESCRT function that would allow us a window of time to assess CTL killing (fig. S4). We used two approaches to block ESCRT function: CRISPR knockout of the *Chmp4b* gene or overexpression of VPS4a^{E228Q} (E228Q, Glu^{228} → Gln), a dominant-negative kinase allele that impairs ESCRT function (fig. S4, A to C) (*10*). We took care to complete our assessment of target killing well in advance of spontaneous target cell death (fig. S4D).

We tested the capacity of OT-I CTLs to kill targets presenting one of four previously characterized peptides that demonstrate a range of potencies at stimulating the OT-I TCR: SIINFEKL (N4), the cognate peptide, and three separate variants (in order of highest to lowest affinity), SIITFEKL (T4), SIIQFEHL (Q4H7), and SIIGFEKL (G4) (*18*, *19*). Target cells were pulsed with peptide, washed, transferred to 96-well plates, and allowed to adhere before the addition of OT-I CTLs. Killing was assessed by monitoring the uptake of a fluorogenic caspase 3/7 indicator (Fig. 4, A to D, and fig. S5A). Killing was significantly more efficient in ESCRT-inhibited target cells for both CRISPR depletion of Chmp4b (Fig. 4, A to C) and expression of the dominant-negative VPS4a^{E228Q} (Fig. 4D). The difference in killing between the ESCRT-inhibited and control cells was greater when the lower-potency T4, Q4H7, and G4 peptides were used. Nevertheless, ESCRT inhibition moderately improved killing efficiency even in the case of the high-potency SIINFEKL peptide. ESCRT inhibition had no effect on MHC class I expression on the surface of target cells (fig. S5B). Thus, ESCRT inhibition could sensitize target cells to perforin- and granzyme-mediated killing, especially at physiologically relevant TCR-peptide MHC affinities.

We next directly tested the effects of ESCRT inhibition when target cells were exposed to both recombinant perforin (Prf) and granzyme B (GZMB), the most potently proapoptotic granzyme in humans and mice (*20*). Prf alone at high concentrations can lyse cells (*4*), so we first determined a sublytic Prf concentration that would temporarily permeabilize the plasma membrane but permit the cells to recover. B16-F10 cells expressing either VPS4a^{WT} (WT, wild-type) or VPS4a^{E228Q} were exposed to a range of Prf concentrations in the presence of PI, and cell viability and PI uptake were assessed using flow cytometry. Cells that expressed dominant-negative VPS4a^{E228Q} were more sensitive to Prf alone than ESCRT-competent cells (Fig. 4, E and F). At 160 ng/ml Prf, there was no significant difference in cell viability for either condition. Cells in the live gate that were PI+ had been permeabilized by Prf but recovered. Although the percentage of PI+ live cells was similar under both sets of conditions, the mean fluorescence intensity of PI was higher in live ESCRT-inhibited cells (fig. S6). A delay in plasma membrane resealing could account for this difference.

We reasoned that delaying perforin pore repair might also enhance GZMB uptake into the target. ESCRT-inhibited cells were more sensitive to combined perforin-GZMB when cell death was measured by Annexin V staining (Fig. 4, G and H). Similar results were observed when these experiments were repeated with a murine lymphoma cancer cell line (fig. S7). The observation that ESCRT-inhibited target cells are more sensitive to both CTL-secreted and Prf-GZMB supports the hypothesis that the ESCRT pathway contributes to membrane repair after Prf exposure.

Escaping cell death is one of the hallmarks of cancer. Our findings suggest that ESCRT-mediated membrane repair of perforin pores may restrict accessibility of the target cytosol to CTL-secreted granzyme, thus promoting survival of cancer-derived cells under cytolytic attack. Although other factors may contribute to setting the threshold for target susceptibility to killing, the role of active repair of perforin pores must now be considered as a clear contributing factor.

REFERENCES AND NOTES

1. S. Halle *et al.*, *Immunity* **44**, 233–245 (2016).
2. B. Weigelin *et al.*, *Nat. Commun.* **12**, 5217 (2021).
3. N. W. Andrews, P. E. Almeida, M. Corrotte, *Trends Cell Biol.* **24**, 734–742 (2014).
4. D. Keefe *et al.*, *Immunity* **23**, 249–262 (2005).
5. J. A. Lopez *et al.*, *Blood* **121**, 2659–2668 (2013).
6. A. J. Jimenez *et al.*, *Science* **343**, 1247136 (2014).
7. L. L. Scheffer *et al.*, *Nat. Commun.* **5**, 5646 (2014).
8. Y.-N. Gong *et al.*, *Cell* **169**, 286–300.e16 (2017).
9. S. Rühl *et al.*, *Science* **362**, 956–960 (2018).
10. K. A. Hogquist *et al.*, *Cell* **76**, 17–27 (1994).
11. G. Knott, H. Marchman, D. Wall, B. Lich, *J. Neurosci.* **28**, 2959–2964 (2008).
12. C. S. Xu *et al.*, *eLife* **6**, e25916 (2017).
13. C. S. Xu *et al.*, *Nature* **599**, 147–151 (2021).
14. D. P. Hoffman *et al.*, *Science* **367**, eaaz5357 (2020).
15. A. T. Ritter *et al.*, *Immunity* **42**, 864–876 (2015).
16. J. C. Stinchcombe, E. Majorovits, G. Bossi, S. Fuller, G. M. Griffiths, *Nature* **443**, 462–465 (2006).
17. I. Voskoboinik, J. C. Whisstock, J. A. Trapani, *Nat. Rev. Immunol.* **15**, 388–400 (2015).
18. M. A. Daniels *et al.*, *Nature* **444**, 724–729 (2006).
19. A. C. Richard *et al.*, *Nat. Immunol.* **19**, 849–858 (2018).
20. D. Kaiserman *et al.*, *J. Cell Biol.* **175**, 619–630 (2006).

ACKNOWLEDGMENTS

We thank members of the Mellman laboratory for advice, discussion, and reagents; B. Haley for assistance with plasmid construct design; the Genentech FACS Core Facility for technical assistance; S. Van Engelenburg of Denver University for invaluable discussions and guidance; A. Wanner, S. Spaar, and the Ariande AI AG (https://ariadne.ai/) for assistance with FIB-SEM segmentation, CLEM coregistration, data presentation, and rendering; D. Bennett of the Janelia Research Campus for assisting with data upload to https://openorganelle.janelia.org; and the Genentech Postdoctoral Program for support.
Funding: A.T.R. and I.M. are funded by Genentech/Roche. C.S.X., G.S., A.W., D.A., N.I., and H.F.H. are funded by the Howard Hughes Medical Institute (HHMI). **Author contributions:** A.T.R. designed and performed experiments and analyzed the data. A.T.R. and I.M. wrote the manuscript with input from all coauthors. G.S. and M.F. performed high-pressure freezing. G.S., D.P.H., and N.A. collected cryo–fluorescence images. G.S. and N.I. performed freeze-substitution and resin embedding. G.S. prepared samples for FIB-SEM. G.S. and C.S.X. conducted FIB-SEM imaging. C.S.X. processed FIB-SEM data. D.A. and A.W. developed FIB-SEM data analysis. D.A. performed FIB-SEM data analysis. A.W. prepared Figs. 1 and 3. I.V. and J.T. provided valuable reagents and consulted on experiments. H.F.H. and I.M. supervised the work and provided support. **Competing interests:** A.T.R. and I.M. declare that they are Genentech/Roche employees. C.S.X. and H.F.H. are the inventors of a US patent assigned to HHMI for the enhanced FIB-SEM systems used in this work ("Enhanced FIB-SEM systems for large-volume 3D imaging," US Patent 10,600,615). The authors declare no other competing interests. **Data and materials availability:** All data are available in the main text or the supplementary materials. Full FIB-SEM datasets and segmentations are freely available to be explored on https://openorganelle.janelia.org, with direct links and data accession numbers provided in the supplementary materials. Cell lines and plasmids are available from I.M. upon request.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abl3855
Materials and Methods
Figs. S1 to S7
References (*21*–*25*)
MDAR Reproducibility Checklist
Movies S1 to S7

9 July 2021; resubmitted 22 December 2021
Accepted 4 March 2022
10.1126/science.abl3855

PROTEIN DESIGN

Computational design of mechanically coupled axle-rotor protein assemblies

A. Courbet[1,2,3]†, J. Hansen[1]†, Y. Hsia[1,2], N. Bethel[1,2,3], Y.-J. Park[1], C. Xu[1,2,3], A. Moyer[1,2], S. E. Boyken[1,2]‡, G. Ueda[1,2], U. Nattermann[1,2], D. Nagarajan[1,2], D. Silva[1,2,4,5], W. Sheffler[1,2], J. Quispe[1], A. Nord[6], N. King[1,2], P. Bradley[7], D. Veesler[1,3], J. Kollman[1], D. Baker[1,2,3]*

Natural molecular machines contain protein components that undergo motion relative to each other. Designing such mechanically constrained nanoscale protein architectures with internal degrees of freedom is an outstanding challenge for computational protein design. Here we explore the de novo construction of protein machinery from designed axle and rotor components with internal cyclic or dihedral symmetry. We find that the axle-rotor systems assemble in vitro and in vivo as designed. Using cryo–electron microscopy, we find that these systems populate conformationally variable relative orientations reflecting the symmetry of the coupled components and the computationally designed interface energy landscape. These mechanical systems with internal degrees of freedom are a step toward the design of genetically encodable nanomachines.

Intricate protein nanomachines in nature have evolved to process energy and information by coupling biochemical free energy to mechanical work. Among the best studied and most sophisticated are protein rotary machines such as the F_1 motor of adenosine triphosphatase or the bacterial flagellum, which contain axle-like and ringlike symmetric protein components capable of constrained dynamic motion relative to each other (*1*–*3*). Feynman's 1959 lecture on nanotechnology as a means to leverage the properties of materials at the molecular scale (*4*) inspired interest in synthetic nanomachines (*5*, *6*). Synthetic chemists were the first to design molecules with mechanically coupled components (*7*–*9*). Nucleic acid nanotechnologies have more recently been used to construct rotary systems (*10*). Designed dynamic protein mechanical systems are of great interest given the richer functionality of proteins, but with this functionality comes more complex folding and a greater diversity of noncovalent interactions, which, despite recent advances in design of static protein nanostructures (*11*–*19*), has made the design of protein machines an outstanding challenge (*20*).

We explored the design of protein mechanical systems through a first-principle, bottom-up approach that focuses on operational concepts independent from the complex evolutionary trajectory of natural nanomachines. Previous two-component protein assembly design studies have focused on nanomaterials such as icosahedral nanocages (*21*) and two-dimensional (2D) arrays (*19*) in which the components have fixed orientations relative to one another. In this work, we sought to design a nanoscale simple machine or kinematic pair (*22*, *23*) in which two mechanically coupled protein components undergo Brownian diffusion along internal degrees of freedom (DOFs). We used a hierarchical design approach with three steps: (i) de novo design of stable and rigid protein components suitable for assembly into constrained mechanical systems, (ii) directed self-assembly of these components into hetero-oligomeric complexes, and (iii) shaping of the multistate energetic landscape along the mechanical DOFs. A major challenge is to design the interface between the two designed rigid bodies to have sufficiently low energy to drive self-assembly, while still allowing relative motion of the components. We started from a machine blueprint that consists of two coupled structural components resembling an axle and rotor (Fig. 1A), in which, similar to natural protein rotary systems, the features of the energy landscape are determined by the symmetry of the interacting components, their shape complementarity, and specific interactions across the interface.

Computational design of protein mechanical components

We first sought to design ringlike protein topologies, or rotors, with a range of inner diameter sizes capable of accommodating an axle-like binding partner (Fig. 1B). In a first approach, we started from de novo α-helical tandem repeat proteins (*24*) and redesigned them to form C1 single-chain structures or symmetric C3 or C4 homo-oligomers. In a second approach, we started from de novo helical repeat proteins (DHRs) and helical bundle heterodimers and used a hierarchical design procedure based on architecture-guided rigid helical fusion (*14*) to build C3 and C5 cyclic symmetric rotor structures. To facilitate subsequent microscopy characterization and modularity, we fused another set of DHRs at the outer side of the rotors, generating armlike extensions (Fig. 1, A and B). Synthetic genes encoding these rotor designs (12xC3s, 12xC4s, 2xC5s) were synthesized and the proteins expressed in *Escherichia coli*. All designed proteins were soluble after purification on nickel-nitrilotriacetic acid (Ni-NTA) columns, and ~23% (6/26) had size exclusion chromatography (SEC) profiles that matched the expected theoretical elution profile for the oligomerization state (figs. S1 and S2 and table S1). These designs were further examined using small-angle x-ray scattering (SAXS) (*25*, *26*), negative stain electron microscopy, or cryo–electron microscopy (cryo-EM) (fig. S1). For the C3_R1 rotor, SAXS data analysis was consistent with the computational model [volatility ratio (Vr) = 4.684] (table S2 and fig. S2), and we were able to determine using cryo-EM a 6.0-Å 3D reconstruction that was close to the design model [backbone root mean square deviation (RMSD) = 3.451 Å] (Fig. 1B; figs. S1, S4, and S5; and table S3). Similar results were obtained for another design of the same topology (C3_R2) (fig. S1). For the C4 design C4_1, we obtained a 7.9-Å cryo-EM density map closely consistent with the design model (backbone RMSD = 1.8 Å) (Fig. 1B; figs. S1, S5, and S6; and table S3). C3 and C5 rotors with larger inner diameter and different topology (C3_R3 and C5_2) were characterized using negative stain EM, yielding low-resolution 3D reconstructions consistent with the design model (Fig. 1B and fig. S1).

We next sought to design high–aspect ratio protein components, or axles, onto which the designed rotor protein could be threaded, using three different design approaches. In a first approach, single-helix backbones were parametrically generated, and the sequence was optimized in D2, D3, or D4 dihedral symmetry using buried hydrogen bond networks and hydrophobic contacts to produce self-assembling homo-oligomer interfaces with the high level of specificity needed for dihedral assembly (Fig. 2A). To increase the total mass and diversify the shape for subsequent EM analysis, the termini of these rod-shaped structures were rigidly fused to cyclic homo-oligomers of matching symmetry (i.e., D*n* dihedral assemblies were fused with C*n* cyclic assemblies) to create dumbbell-shaped structures. In a second approach, two copies of designed cyclic homo-oligomers were assembled into dihedral structures by connecting them with rigid helical bundle connectors built using fragment sampling (Fig. 2B). In a third approach, parametrically generated homotrimer backbones consisting of helical hairpin monomer topologies (*27*) were circularly permuted and

[1]Department of Biochemistry, University of Washington, Seattle, WA, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA, USA. [3]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [4]Division of Life Science, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. [5]Monod Bio, Inc., Seattle, WA, USA. [6]Centre de Biologie Structurale (CBS), INSERM, CNRS, Université Montpellier, Montpellier, France. [7]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA.
*Corresponding author. Email: dabaker@uw.edu
†These authors contributed equally to this work.
‡Present address: Outpace Bio, Inc., Seattle, WA, USA.

elongated to generate extended C3 homo-oligomers (Fig. 2C). Details of the methods, as well as scripts for carrying out the design calculations, are provided in the supplementary materials. Synthetic genes encoding axle designs generated from the three approaches (12xC3s, 12xC5s, 12xC8s, 6xD2s, 12xD3s, 6xD4s, 6xD5s, and 12xD8s) were obtained, and the proteins were expressed in *E. coli*. The designed proteins that were well expressed, soluble, and readily purified by Ni-NTA affinity chromatography were further purified on SEC. Success rates for the first, second, and third approach were 37.5% (6/16), 43% (14/32), and 33% (4/12), respectively, as assessed by the match between estimated molecular weight (MW) from SEC with the MW of the design model (Fig. 2D; figs. S1 to S3; and table S1). Designs with matching SEC traces were further examined using SAXS, negative stain EM, and cryo-EM (figs. S1 to S3).

The first approach generated D2, D3, and D4 axle-like structures with folds featuring interdigitated helices with extended hydrogen bond networks. We obtained a 4.2-Å 3D reconstruction of a D3 axle (D3_3) with backbone nearly identical to the design model (backbone RMSD = 1.9 Å) (Fig. 2A and figs. S3, S4, and S7); SAXS data were also consistent with the design model (Vr = 6.0) (table S2 and figs. S1 and S2). The central homohexameric 50-residue helices (D3_2) could also be solubly expressed and formed an oligomeric self-assembly that eluted at the expected volume (fig. S3 and table S1). D3 design D3_1, consisting of 36-residue-long single helices, was produced by chemical peptide synthesis and assembled into a homohexamer (figs. S3 and S8), and fusion to wheel-like C3s generated a larger D3 oligomer as designed (D3_4) (fig. S3). A D4 peptide homo-oligomer designed using the same approach (D4_1) had a SEC profile consistent with the expected oligomeric state (figs. S2 and S3 and table S1). Negative stain EM of a D2 design (D2_2) yielded a low-resolution 3D reconstruction with the overall features of the design model (Fig. 2D and fig. S3); the corresponding central 50-residue D2 peptide (D2_1) could also be expressed, and the SEC elution volume corresponded to the expected oligomeric state (fig. S3 and table S1).

The second approach generated D3, D4, D5, and D8 axle-like structures, with interdigitated helices with internal cavities in the D5 and D8 cases where each central helix only contacts the two neighboring ones (Fig. 2B). We obtained a 7.4 -Å electron density map of a D8 design (D8_1) revealing a backbone structure nearly identical to the design model (backbone RMSD = 2.9 Å) (Fig. 2B and figs. S3, S5, and S6). This cylinder-shaped homodecahexamer has a large central cavity, an 84-residue helix, and opposing N and C termini close to its center (Fig. 2B and fig. S3). Negative stain EM 3D reconstructions of D8_2, D8_3, D5_2, and D4_2 were consistent with the design models (Fig. 2D and fig. S3). We converted several of these designs from dihedral to cyclic symmetry by connecting N and C termini, and two such designs, one C5 (C5_1) and one C8 (C8_1), yielded EM reconstructions with good agreement with the design model (Fig. 2D and figs. S1 and S3). SAXS profiles of additional designs (4xD3s, 2xD4s, and 1xD5) were consistent with the design models with Vr < 10 in most cases and measured MW within 15% of the design model for D3_1 and D3_8, and within 1% for D5_1 (table S2 and figs. S2 and S3).

The third approach yielded four C3 axles with smaller aspect ratios and overall sizes, containing a large wheel-like feature at one end, a narrow central three-helix section, and a six-helix section at the other end. SAXS profiles together with SEC traces suggested that the designed oligomerization state is realized in solution (Vr ~ 12) (tables S1 and S2 and figs. S1 and S2). For design C3_A1, we obtained a low-resolution cryo-EM map that recapitulates the general features of the design model, with prominent C3 symmetric DHR extremities and opposing prism-like extensions (Fig. 2C and figs. S1 and S4).

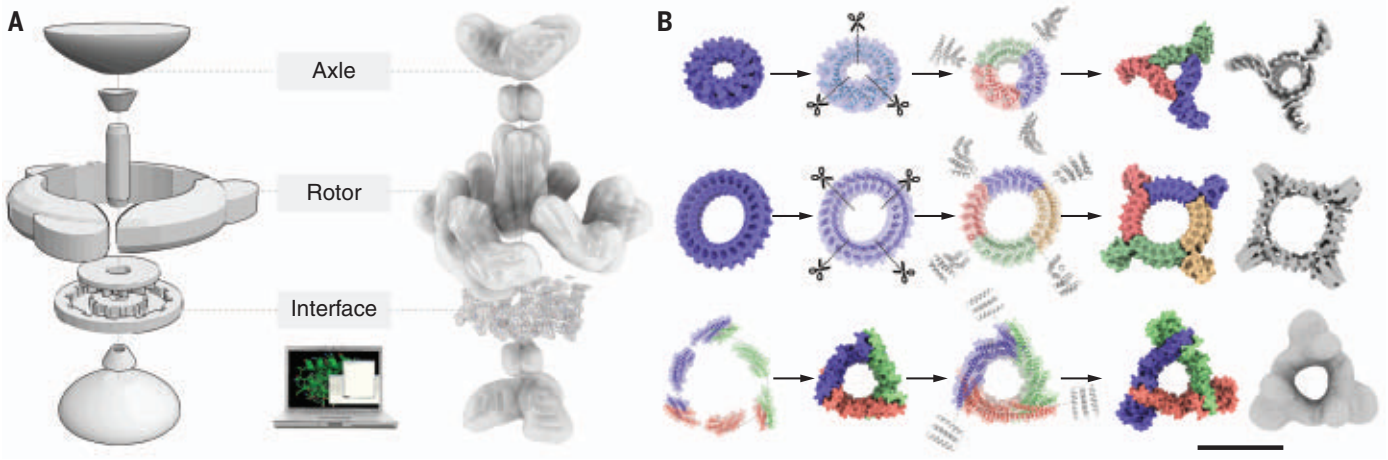


Fig. 1. Overview of protein machine assembly and rotor component design approaches. (**A**) (Left) A blueprint of a simple two-component machine consisting of an assembly of an axle and a rotor mechanically constrained by the shape of the interface between the two. (Middle) Systematic generation by computational design of a structurally diverse library of machine components and design of interfaces between axle and rotor that mechanically couple the components and direct assembly. (Right) Example of hierarchical design and assembly of a protein machine from axle and rotor components, here a D3 axle and C3 rotor, and interacting interface residues. Wheel-like cyclic DHRs are fused to the end of the axle and rotor components to increase mass and provide a modular handle and a structural signature to monitor conformational variability. (**B**) Hierarchical design strategies for rotor components. (Top) A single-chain C1 symmetric and internally C12 symmetric α-helical tandem repeat protein is split into three subunits, and each is fused to DHRs by means of helical fusion (HelixFuse) to generate a C3 rotor (C3_R1) with an internal diameter of 28 Å. The 6.0-Å cryo-EM electron density (shown in gray) shows agreement with the design model. (Middle) A single-chain C1 symmetric and internally C24 symmetric α-helical tandem repeat protein is split into four subunits, and each is fused to DHRs to generate a C4 rotor (C4_1) with an internal diameter of 57 Å. The 7.9-Å cryo-EM electron density (shown in gray) shows agreement with the design model. (Bottom) Hetero-oligomeric helical bundles and DHRs are fused and assembled into a higher-ordered closed C3 structure through helical fusion, after which another round of helical fusion protocol is used to fuse DHRs to each subunit, to generate a C3 rotor (C3_R3) with an internal diameter of 41 Å. The negative stain electron density (shown in gray) shows agreement with the design model. Monomer subunits are colored by chain. Scale bar, 10 nm.

Design of mechanically coupled axle-rotor assemblies

We next investigated the construction of mechanically constrained axle-rotor assemblies from the designed axles and rotors. As noted above, an inherent challenge for the de novo design of dynamic protein complexes is to incorporate sufficient energetically favorable interactions to enable directed self-assembly without creating deep energy minima that lock the assembly into a single state and prevent Brownian diffusion along the mechanical DOFs. We explored three approaches for constructing axle-rotor assemblies, which result in interfaces with widely varying energetics, shape complementarity, and symmetry.

First, we sought to construct two-component assemblies in which the rigid body orientation of the axle and rotor was minimally constrained. We designed symmetry-mismatched axle-rotor interfaces with low orientational specificity and loose interface packing, allowing only small numbers of close contacts across the interface and using primarily electrostatic interactions between rotor and axle, which are longer range and less dependent on shape matching than the hydrophobic interactions

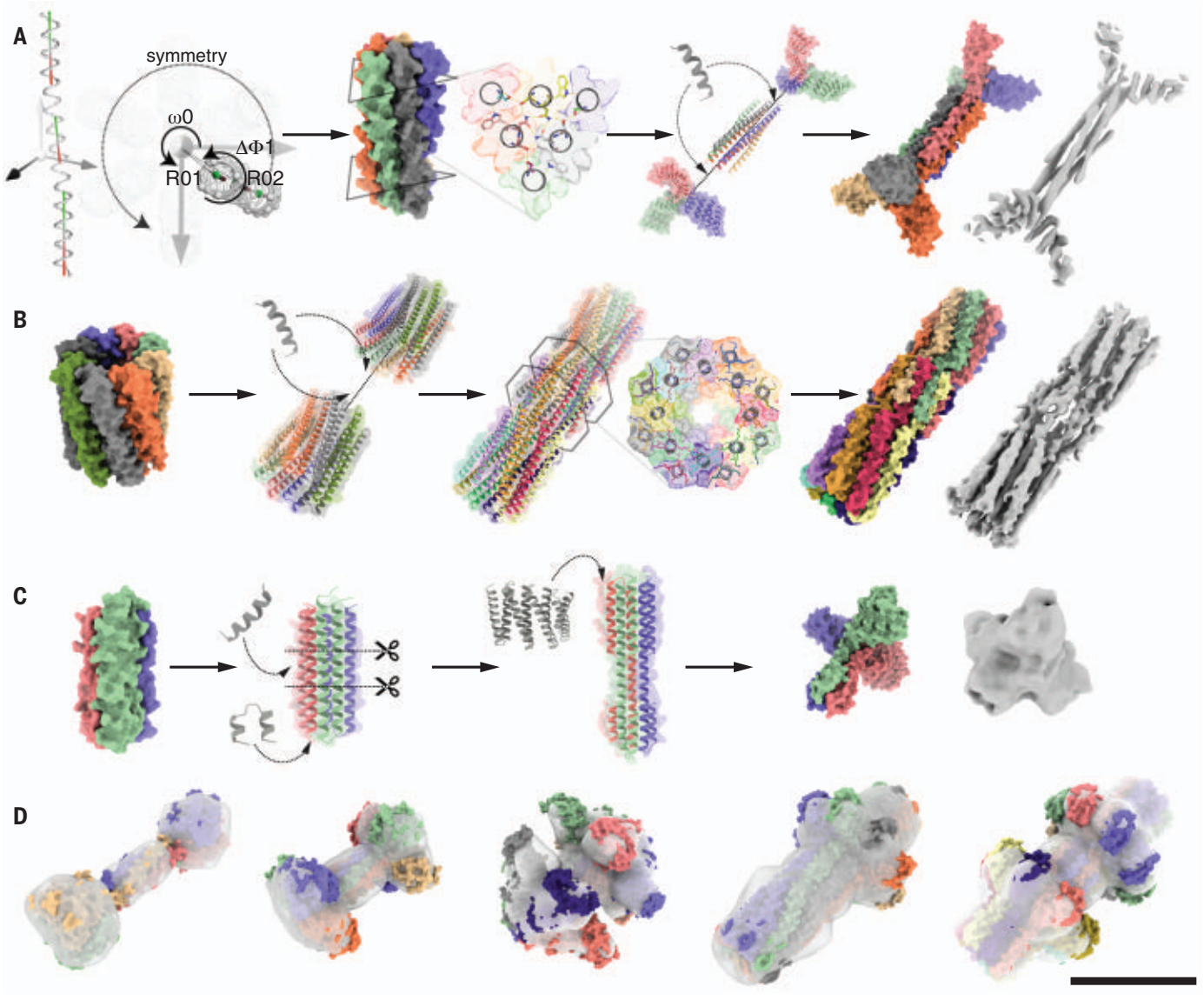


Fig. 2. Design of axle machine components. (**A**) Hierarchical design of a D3 symmetric homohexamer axle (D3_3). Parametric design of interdigitated helices in D3 symmetry was achieved by sampling supercoil radius (R01, R02), helical phase ($\Delta\varphi_{1\text{-}1}$, $\Delta\varphi_{1\text{-}2}$), supercoil phase ($\Delta\varphi_{0\text{-}1}$, $\Delta\varphi_{0\text{-}2}$) of two helical fragments, and the *z*-offset and supercoil twist (ω_0). The interface was designed using the HBNet protocol to identify hydrogen bond networks spanning the six helices mediating high-order specificity. The design was then fused to C3 wheel-like homotrimers using RosettaRemodel. The 4.2-Å cryo-EM electron density is consistent with the design model. (**B**) Hierarchical design of a D8 axle (D8_1). Interdigitated helical extensions at the termini of a parametrically designed C8 homohexamer were sampled using Rosetta BluePrintBuilder, and hydrogen bond networks were identified using HBNet, while sampling rotation and translation in D8 symmetry using Rosetta SymDofMover. The 7.4-Å cryo-EM electron density is in close agreement with the design model. (**C**) Hierarchical design of a C3 homotrimer axle (C3_A1). A parametrically designed C3 homotrimer was circularly permutated and an extra heptad repeat added to increase the aspect ratio, after DHRs were fused to each subunit using HelixFuse. The negative stain electron density is consistent with the design model. (**D**) Additional axle components overlaid with experimental negative stain electron density, corresponding to (from left to right) D2 (D2_2), D4 (D4_2), D5 (D5_2), C8 (C8_1), and D8 (D8_3) designs. Model monomer subunits are colored by chain, and electron densities are shown as gray surfaces. Scale bar, 10 nm.

generally used in protein design. To prevent potential disassembly at low concentrations owing to weak axle-rotor interactions, we sought to kinetically trap the rotor around the axle by installing disulfide bonds at the rotor subunit-subunit interfaces. To gain stepwise control on the in vitro assembly process, we introduced buried histidine-mediated hydrogen bond networks at the asymmetric interfaces between rotor subunits to enable pH-controlled rotor assembly (fig. S9, and see methods in the supplementary materials). To test this approach, we selected three of the machine components described above—a D3 axle, a C3 rotor, and a C5 rotor—and constructed axle-rotor assemblies with D3-C3 and D3-C5 symmetries (designs D3-C5 and D3-C3, respectively) (Fig. 3A and fig. S10). To thread axles and rotor together, we computationally sampled rotational and translational DOFs and designed complementary electrostatic interacting surfaces excluding positively charged residues on the axle (lysine and arginine) and negatively charged residues on the rotor (aspartate and glutamate). Given the shape complementarity between the internal diameter of the rotors and the axle thickness, the interface is tight for D3-C3, constraining the rotor on the axle, and loose for D3-C5. By design, the D3-C3 can rotate and translate along the main symmetry axis (z), whereas the D3-C5 rotor has rotation and translation components along x, y, and z (Fig. 3, A and B, and fig. S11). Synthetic genes encoding the one axle and two rotor designs were obtained, and the proteins were separately expressed in *E. coli* and purified by Ni-NTA affinity chromatography and SEC, which indicated that the surface redesign did not affect solubility or oligomerization state (figs. S1 and S3). After stoichiometric mixing of the designed D3 axle and C3 rotor, EM analysis showed a collection of assembled and isolated axle and rotor molecules (fig. S9A, top panel). After dropping the pH and reducing the disulfide, the particles appeared as a mixture of opened, linear, and hard-to-distinguish particles (fig. S9A, middle panel). After restoring the pH under oxidizing conditions, the particles appeared fully assembled by EM (fig. S9A, bottom panel). Biolayer interferometry assays showed that the rotor and axle associated rapidly with an approximate association rate of $10^3\ M^{-1}{\cdot}s^{-1}$ and a dissociation constant (K_d) in the micromolar range (fig. S12). Similar results

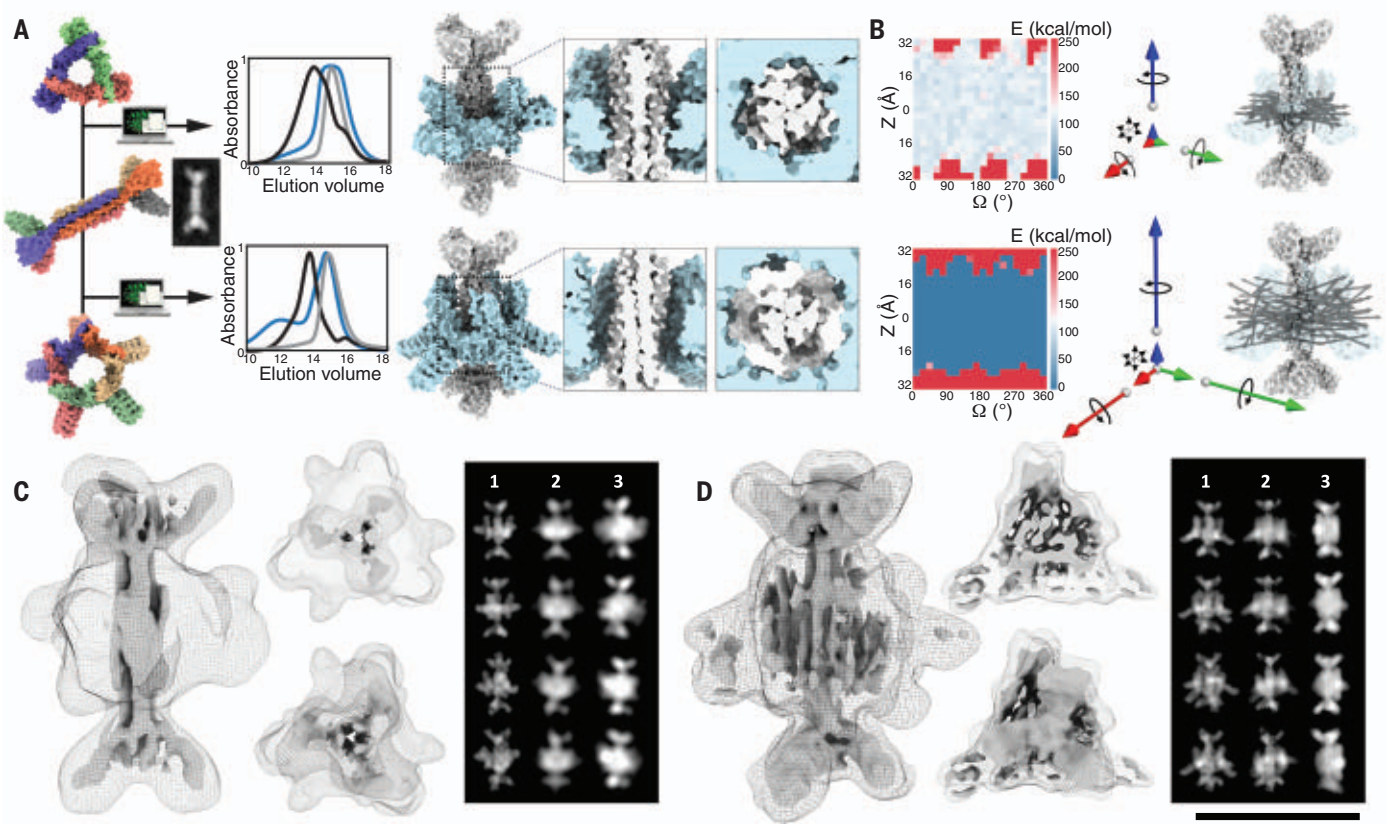


Fig. 3. Design of symmetry-mismatched D3-C3 and D3-C5 axle-rotor assemblies. (**A**) (Left to right) Models of a D3 axle (D3_3) and C3 (C3_R3) and C5 (C5_2) rotors and cryo-EM 2D average of axle alone before assembly. Overlaid SEC chromatograms (absorbance at 215 nm) of axle (gray), rotor (blue), and full assembly (black). Models of D3-C3 and D3-C5 assemblies with top-view and side-view close-up on interfaces; shape and symmetry results in different DOFs. (**B**) (Left) 2D rotation-translation energy landscapes showing a large area of low energy where the rotor can be positioned on the axle. (Right) MD simulation results are shown as vectors whose magnitude corresponds to the computed mean square displacement of the rotor relative to the axle along the six DOFs. The D3-C3 system is largely constrained to rotation along the z axis (blue), whereas the D3-C5 assembly allows rotation along x (green), y (red), and z and translation in z, x, and y. N- and C-terminal unit vectors of an ensemble of MD trajectories are superimposed on an axle-rotor model structure. (**C**) (Left) 3D cryo-EM reconstruction of D3-C3, processed in D3 at 7.8-Å resolution suggests that the rotor sits midway across the D3 axle, consistent with the designed mechanical DOF. The maps are shown in side views, end-on views, and transverse slices, as surface for the axle and as mesh for the rotor, at two different thresholds. (Right) Simulated 2D class averages without (1) and with (2) conformational variability, and experimental averages (3). (**D**) (Left) 3D cryo-EM reconstruction of D3-C5, processed in C1 at 8.6-Å resolution, has the overall features of the designed structure, shown as surface and mesh at different thresholds. The 2D averages capture secondary structure corresponding to the C5 rotor but could not be fully resolved, which is consistent with the rotor populating conformationally variable states. (Right) Simulated 2D class averages without (1) and with (2) conformational variability, and experimental averages (3). Scale bar for cryo-EM density, 10 nm.

were obtained with D3-C5 rotary assemblies, and SEC and SAXS profiles were in agreement with the design model in both cases (Vr < 15) (tables S1 and S2 and figs. S2 and S10).

Second, we experimented with more direct steric coupling to limit conformational variability primarily to rotation of the rotor around the axle. We used shape-complementary axle and rotor components to enable the incorporation of steric constraints restricting translation, leveraging Rosetta's ability to design tightly packed interfaces and hydrogen bond network–mediated specificity (*27*). We designed seven axle-rotor assemblies using this approach: three with C3 symmetric axles with C1 rotors (C3-C1_1 to C3-C1_3) (fig. S10) and four larger designs with C3 axles and rotors (C3-C3_1 to C3-C3_4) (Fig. 4A and fig. S10) with DHR arm extensions. The C3 symmetry matching of the rotor and axle differs from the mismatching in the other designed assemblies, and the extent of

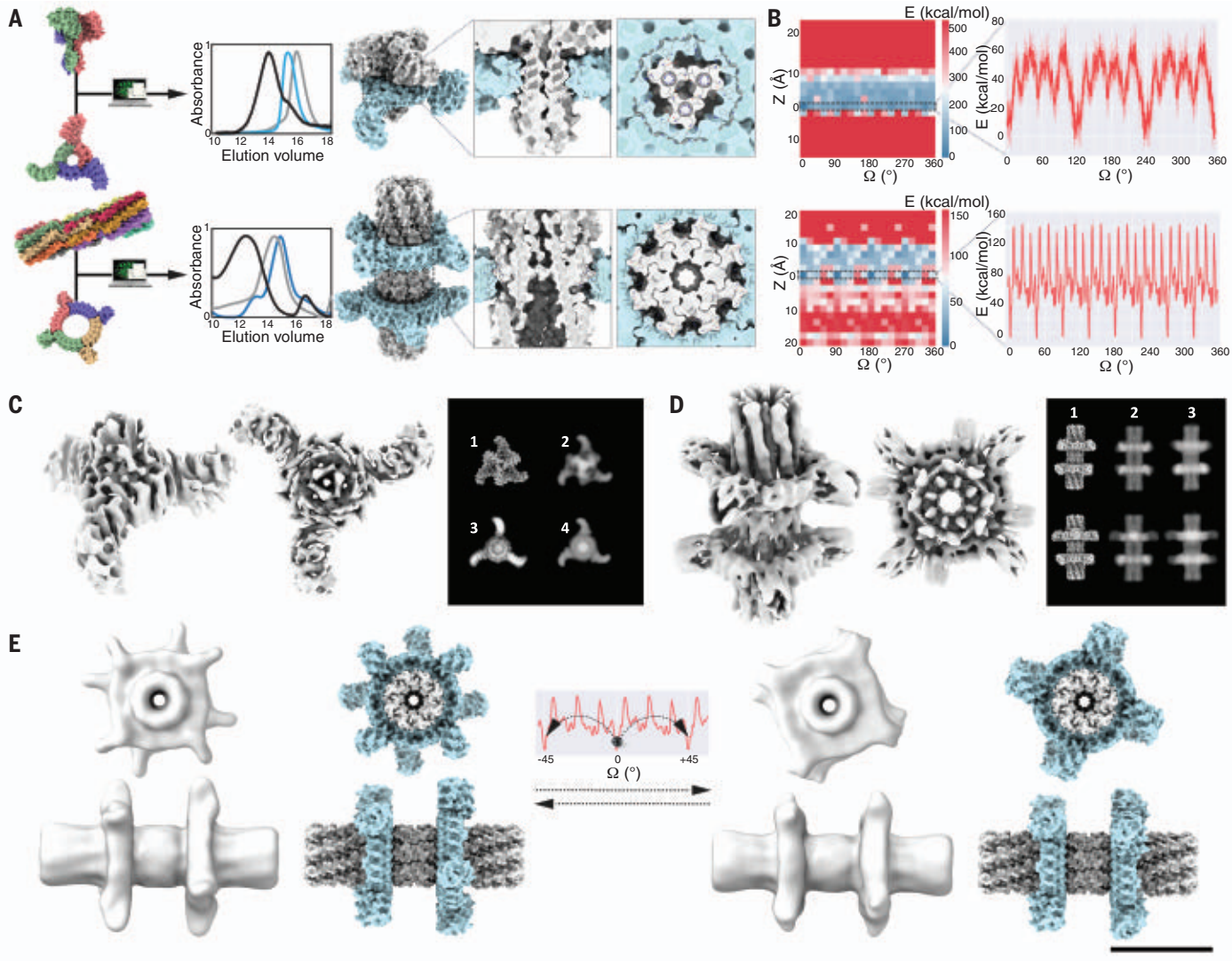


Fig. 4. Computational sculpting of the energy landscape by design of interface side-chain interactions. (**A**) (Left to right) Models of C3 axle (C3_A1), C3 rotor (C3_R1), D8 axle (D8_1), and C4 rotor (C4_1) used to design C3-C3 and D8-C4 axle-rotor assemblies. Overlaid SEC chromatograms (absorbance at 215 nm) of axle (gray), rotor (blue), and full assembly (black). Models of symmetry-matched C3-C3_1 and quasi-symmetric D8-C4 assemblies and close-ups on the interface reveal the shape-complementary cogwheel topology. (**B**) Energy landscapes corresponding to the C3-C3 (top) and D8-C4 (bottom) axle-rotor assemblies. (Left) 2D rotation-translation energy landscapes showing a narrow band of low energy where the rotor sits on the axle. (Right) 1D rotational energy landscape has three main minima corresponding to the C3 symmetry of the interface with nine additional lesser energy minima for C3-C3 and eight main energy minima corresponding to the C8 symmetry of the interface and additional 18 lesser minima for D8-C4. The energy landscapes were computed by scoring 10 independent Rosetta backbone and side-chains relax and minimization trajectories (solid red line with error bars depicting the standard deviation; kilocalories per mole, as calculated by Rosetta). (**C**) Single-particle cryo-EM analysis of the C3-C3 assembly. The rotor is evident in the 6.5-Å resolution electron density in the side and top views; only a portion of the axle is resolved. In the panel to the right, the experimental 2D class averages (3) match the projection of the design model (1) more closely with conformational variability (4) than without (2). (**D**) Single-particle cryo-EM analysis of the designed D8-C4 rotor. The electron density (in gray) at 5.9-Å resolution shows the main features of the designed structure and two distinct rotational states (1), also visible in the simulated projections (2), which closely resemble the experimental 2D class average (3). (**E**) 3D variability analysis and computed rotational landscape of the D8-C4 axle-rotor assembly. The two resolved structures (shown in gray on the left and right) are separated by a 45° rotational step. Corresponding computational models are shown in spacefill (blue and gray). Top row: top view; bottom row: side view. Scale bar, 10 nm.

alignment of axle and rotor DHR arms relative to each other provides a measure of conformational variability. Design was carried out by systematically sampling rotational and translational DOFs, removing arrangements with backbone-to-backbone clashes (see methods), and then using the Rosetta HBNet protocol and FastDesign (*28*) to identify interacting residues and optimize the interface energy. Each interface design trajectory generates widely different periodic energy landscapes according to interface metrics and design specifications (fig. S13) and results in shape-complementary axle-rotor interfaces with an overall cogwheel topology. C3-C1 designs were experimentally screened for assembly by expressing rotor and axle pairs bicistronically and carrying out Ni-NTA purification relying on a single His tag on the rotor component (fig. S14A). Six out of 12 designs expressed solubly and copurified, suggesting that the two components assembled in cells (fig. S14B), and three designs (C3-C1_1 to C3-C1_3) (fig. S10) were selected for further characterization. The SEC profiles in combination with native mass spectrometry indicated an oligomeric state consistent with the designed assembly, and SAXS data were also consistent with the design model (Vr < 12 and MW within ~10% of expected values for C3-C1_3, and ~15% for C3-C1_1 and C3-C1_2) (tables S1 and S2 and figs. S2, S10, and S14, C and D). The C3-C3 designs (C3-C3_1 to C3-C3_4) (fig. S10) were screened for in vitro assembly by stoichiometric mixing of axle and rotor, followed by SEC and SAXS analysis, which were consistent with assemblies of the expected oligomeric state (Vr < 10) (tables S1 and S2 and figs. S2 and S10). Biolayer interferometry showed that the designed C3 axle and C3 rotor rapidly assemble with an approximate association rate of 10^3 $M^{-1}{\cdot}s^{-1}$ and a K_d in the micromolar range (fig. S12).

Third, we sought to design further constrained axle-rotor assemblies by increasing the surface area of the interfaces between axle and rotor to enable more extensive sculpting of the energy landscape. We designed a symmetry-mismatched assembly consisting of a D8 axle around which two C4 rotors are assembled (D8-C4), a symmetry-mismatched assembly consisting of a C5 axle and C3 rotor (C5-C3_1 and C5-C3_2), and a C8-C4 assembly corresponding to a circular permutation version of D8-C4 (C8-C4) (Fig. 4A and fig. S10). The D8-C4 assembly with one axle for two rotors tests the incorporation of multiple coupled rotational DOFs in a multicomponent system and also provides a simple way to monitor the position of rotors relative to each other by experimental structural characterization. For the D8-C4, C5-C3, and C8-C4 designs, because the symmetry of the rotor is internally mismatched to the axle, we used a quasi-symmetric design protocol (see methods). The C4 rotor has internal C24 symmetry, and hence is symmetry-matched to both D8 and C8 axles. In contrast, the C5-C3 arrangement has broken symmetry with a resulting energy landscape with 15 energy minima, with periodicities reflecting the constituent C5 and C3 symmetries (fig. S13). Twelve D8-C4, twelve C5-C3, and six C8-C4 designs were screened for in vitro assembly by isolating axle and rotors individually by Ni-NTA purification and mixed stoichiometrically. We selected four of these designs for further experimental investigation and obtained SEC data indicative of assembly of axle-rotor complexes, while SAXS analysis of a C5-C3 design suggested assembly of the axle-rotor complex (Vr = 6.9 and predicted MW within 6% of expectation) (tables S1 and S2 and figs. S2 and S9). Biolayer interferometry binding kinetics and negative stain EM data were also consistent with quantitative assembly into the designed hetero-oligomeric complex (figs. S10 and S13).

Correspondence between designed energy landscape and observed mechanical DOFs

We subjected one construct from each design approach and symmetry class to single-particle cryo-EM examination and related these data to energy landscape calculations based on the design model (Figs. 3 and 4). Comparison of the electron density data on the axle-rotor assemblies to data on the isolated rotors and axles suggests considerable variation in their rigid body orientations, as summarized in figs. S17 to S19 and S21.

For the D3-C3 and D3-C5 assemblies produced by the first approach, we obtained 2D class averages that clearly resembled projection maps computed from the design models and 3D reconstructions in close agreement with the overall design model topology and designed hetero-oligomeric state (Fig. 3, C and D; figs. S15 and S16; and table S3). For both designs, the D3 axle was clearly visible, and we obtained a high-resolution structure of the axle nearly identical to the design model. 3D reconstructions in C1, C3, and D3 of the D3-C3 axle-rotor assembly at 7.8-Å resolution showed visible density corresponding to the rotor in the middle of the axle with the C3 rotor arms clearly evident (Fig. 3C and fig. S15). 3D reconstructions of the D3-C5 design also showed clear density for the rotor, which could be isolated by masking the axle, but its resolution could not be further improved, as the secondary structure placement relative to the axle appeared variable (Fig. 3D and fig. S16). The particle alignment algorithm is likely dominated by features of the axle that is mostly in side-view in the data (figs. S17 and S18), and thus the lack of resolution of the electron density corresponding to the rotor (see supplementary materials) is probably due to variability in the axle-rotor rigid body transform. CryoSPARC 3D variability analysis (*29*) suggests that the rotor can populate multiple translational and rotational conformational states around the axle (movies S1 to S4). Inspection of the cryo-EM 3D reconstruction also suggests that the rotor arms populate multiple positions along the rotational axis (Fig. 3, C and D, and figs. S17 and S18). Rosetta energy landscapes generated by rotating and translating the rotor relative to the axle suggest that a broad range of orientations are energetically accessible (Fig. 3B), and the rotor-axle rigid body orientation fluctuated in molecular dynamics (MD) simulations, with the D3-C5 assembly showing increased displacement compared with the D3-C3 assembly (Fig. 3B and figs. S11, S17, and S18). Explicit modeling of conformational variability along the designed DOFs was necessary to produce computed projections closely resembling the experimental 2D class averages (Fig. 3, C and D, and fig. S18). Taken together, the cryo-EM data, Rosetta models, and molecular dynamics simulations are consistent with the design goal of constrained mechanical coupling of axle and rotor components (see figs. S17 and S18 for summary of data indicating conformational sampling of rotor-axle rigid body DOFs).

Among the assemblies generated with the second approach, single-particle cryo-EM analysis of a C3-C3 assembly yielded 2D class averages with the axle and rotor clearly visible. Resolution was limited by the orientation bias of the particle in ice, resulting in few side views, but we were able to obtain a 6.5-Å 3D reconstruction that resembled the design model (Fig. 4A; figs. S4, S10, and S19; and table S3). 2D averages and the 3D reconstruction clearly capture the rotor component, but the axle was only partially resolved; the rotor has a mass greater than the C3 axle and clear armlike features, which likely bias the alignment algorithm in its favor. Aligning on the rotor yielded a density map with diffuse density for the axle near the rotor (fig. S19). The contrast between the diffuse density for the axle and the well-resolved density of the rotor likely reflects conformational variability (Fig. 4, C and D, and figs. S4, S18, and S19). The Rosetta energy landscape suggests that the axle-rotor assembly can primarily sample rotational rather than translational DOFs (Fig. 4B), and rotational averaging increased the similarity between projections of the design model and the experimental data (Fig. 4, C and D, and figs. S18 and S19). Taken together, the data (summarized in fig. S19) are consistent with variability along the rotational DOF, in accordance with the designed energy landscape, which has three energy minima at a 60° rotation distance and nine other 30°-spaced degenerate alternative wells separated by low energy barriers (Fig. 4B and figs. S13 and S20).

The D8-C4 design generated by the third approach has a rugged energy landscape, with a dynamic range of 151 kcal/mol (as estimated by Rosetta), and eight steep wells spaced 45° stepwise along the rotational axis corresponding to the high symmetry of the interface (Fig. 4B). Consistent with the deep minima in this landscape, we obtained a cryo-EM map of ~5.9-Å resolution that is close to the design model (Fig. 4, C and D; fig. S6; and table S3). 3D variability analysis calculations using cryoSPARC (*30*) suggested two nearly equiprobable states in which the rotor arms are either aligned or offset, as in the eclipsed and staggered arrangements of ethane (Fig. 4, D and E; figs. S6, S18, and S21; and movie S5). The two rotational states of one rotor relative to the other suggest energy minima spaced by 45° along the rotational axis, consistent with an eightfold step-like feature in the frequency spectrum analysis of the computed energy landscape (fig. S21). Although cryo-EM provides a frozen snapshot of molecules and not a real-time measurement of diffusion, these data (summarized in fig. S20) suggest that the system populates multiple rotational states consistent with the designed energy landscape. Taken together, these results suggest that the explicit design of side-chain interactions and deep energy minima reduces the degeneracy of conformational states observed with purely electrostatic interactions and support a correspondence between the energy landscape and the observed conformational variability.

Conclusions

Our proof-of-concept axle-rotor assemblies demonstrate that protein nanostructures with internal mechanical constraints can now be systematically designed. Key to this advance is the ability to computationally design protein components with complex complementary shapes, symmetries, and topologies, such as the high–aspect ratio dihedral axle structures (D2 homotetramers to D8 homo-16-mers (Figs. 1 and 2) with oligomerization states and sizes considerably larger than previously designed dihedral structures. Our studies of assembly of these shape-complementary homo-oligomeric components into higher-order hetero-oligomeric structures with internal DOFs provide insights toward the design of complex protein nanomachines. First, computational sculpting of the interface between the components can be used to promote self-assembly of constrained systems with chosen internal DOFs. Second, the shape and periodicity of the resulting energy landscape is determined by the symmetry of components, the shape complementarity of the interface, and the balance between hydrophobic packing and conformationally promiscuous electrostatic interactions (Figs. 3, A and B, and 4, A and B). Symmetry mismatch generates assemblies with larger numbers of energy minima than symmetry-matched ones evident in the frequency domain (figs. S13 and S20), and explicit design of close side-chain packing across the interface results in deeper minima and higher barriers than nonspecific interactions (Figs. 3 and 4 and fig. S13). In general, the surface area of the interface between axle and rotor scales with the number of subunits in the symmetry, with larger surface areas providing a larger energetic dynamic range accessible for design (Figs. 3 and 4 and fig. S13). The combination of the conformational variability apparent in the cryo-EM data of D3-C3, D3-C5, and C3-C3 designs (Figs. 3, C and D, and 4, C and D, and figs. S4 and S15 to S19), the Rosetta and MD simulations (Figs. 3B and 4B and fig. S11), and the discrete states observed for the D8-C4 design (Fig. 4, D and E, and figs. S6 and S21) suggests that these assemblies populate multiple rotational states (the axle-rotor assemblies also have multiple symmetrically identical yet physically distinct rotational states—for example, rotation of the C3 rotor around the C3 axle by 120°—which cannot be distinguished by cryo-EM). Our cryo-EM analysis cannot distinguish whether the conformational variability reflects rotational motion or states captured during axle-rotor assembly and does not report on energy barrier heights; time-resolved microscopy at the single-molecule level will be required to reveal the dynamics of transitions between the different states and to relate the computational sculpting of the rotational energy landscapes to Brownian dynamics.

The internal periodic but asymmetric rotational energy landscapes of our mechanically coupled axle-rotor systems provide one of two needed elements for a directional motor. Coupling to an energy input to break detailed balance and drive directional motion remains to be designed: for example, the interface between machine components could be designed for binding and catalysis of a small-molecule fuel (*22*). Symmetry mismatch, which plays a crucial role in torque generation in natural motors (*31*, *32*), can be incorporated in synthetic protein motors, as illustrated here for our axle-rotor assemblies. Modular assembly could lead to compound machines for advanced operation or integration within nanomaterials, and the components can be further functionalized using reversible heterodimer extensions (*33*) (fig. S22). Our protein systems can be genetically encoded for multicomponent self-assembly within cells (fig. S14) or in vitro (figs. S9 and S12). Taken together, these approaches could enable the engineering of a range of nanodevices for medicine, material sciences, or industrial bioprocesses. More fundamentally, de novo design provides a bottom-up platform to explore the fundamental principles and mechanisms underlying nanomachine function that complements long-standing studies of the elaborate molecular machines produced by natural evolution.

REFERENCES AND NOTES

1. W. Junge, N. Nelson, *Annu. Rev. Biochem.* **84**, 631–657 (2015).
2. D. Okuno, R. Iino, H. Noji, *J. Biochem.* **149**, 655–664 (2011).
3. J. C. Deme *et al.*, *Nat. Microbiol.* **5**, 1553–1564 (2020).
4. R. P. Feynman, *Eng. Sci.* **23**, 22–36 (1960).
5. L. Zhang, V. Marcos, D. A. Leigh, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 9397–9404 (2018).
6. K. Drexler, *Trends Biotechnol.* **17**, 5–7 (1999).
7. B. L. Feringa, *J. Org. Chem.* **72**, 6635–6652 (2007).
8. J.-P. Sauvage, *Angew. Chem. Int. Ed.* **56**, 11080–11093 (2017).
9. C. Cheng, J. F. Stoddart, *ChemPhysChem* **17**, 1780–1793 (2016).
10. H. Ramezani, H. Dietz, *Nat. Rev. Genet.* **21**, 5–26 (2020).
11. D. Baker, *Protein Sci.* **28**, 678–683 (2019).
12. G. L. Butterfield *et al.*, *Nature* **552**, 415–420 (2017).
13. Z. Chen *et al.*, *Science* **368**, 78–84 (2020).
14. Y. Hsia *et al.*, *Nat. Commun.* **12**, 2294 (2021).
15. R. A. Langan *et al.*, *Nature* **572**, 205–210 (2019).
16. G. Ueda *et al.*, *eLife* **9**, e57659 (2020).
17. C. Xu *et al.*, *Nature* **585**, 129–134 (2020).
18. R. Divine *et al.*, *Science* **372**, eabd9994 (2021).
19. A. J. Ben-Sasson *et al.*, *Nature* **589**, 468–473 (2021).
20. B. Kuhlman, P. Bradley, *Nat. Rev. Mol. Cell Biol.* **20**, 681–697 (2019).
21. J. B. Bale *et al.*, *Science* **353**, 389–394 (2016).
22. H. Flechsig, A. S. Mikhailov, *J. R. Soc. Interface* **16**, 20190244 (2019).
23. T. A. Hearson, *Philos. Trans. R. Soc. London Ser. A* **187**, 15–40 (1896).
24. L. Doyle *et al.*, *Nature* **528**, 585–588 (2015).
25. K. N. Dyer *et al.*, *Methods Mol. Biol.* **1091**, 245–258 (2014).
26. G. L. Hura *et al.*, *Nat. Methods* **10**, 453–454 (2013).
27. S. E. Boyken *et al.*, *Science* **352**, 680–687 (2016).
28. J. B. Maguire *et al.*, *Proteins* **89**, 436–449 (2021).
29. A. Punjani, D. J. Fleet, *J. Struct. Biol.* **213**, 107702 (2021).
30. A. Punjani, J. L. Rubinstein, D. J. Fleet, M. A. Brubaker, *Nat. Methods* **14**, 290–296 (2017).
31. M. Sobti *et al.*, *Nat. Commun.* **11**, 2615 (2020).
32. D. D. Majewski *et al.*, *Nat. Commun.* **10**, 626 (2019).
33. D. D. Sahtoe *et al.*, *Science* **375**, eabj7662 (2022).
34. A. Courbet, alexiscourbet/Computational-design-of-protein-nanomachines: v1.0, Zenodo (2022); https://doi.org/10.5281/zenodo.6370776.

ACKNOWLEDGMENTS

We thank B. Carragher, C. Potter, E. Eng, L. Yen, and M. Kopylov of the New York Structural Biology Center for assistance and helpful discussions. We especially thank S. Scheres of MRC-LMB for helpful discussions and guidance regarding cryo-EM data processing. We thank the Rosetta@Home user base for donating their computational hours to run forward folding simulations. We thank F. Busch and V. Wysocki at Ohio State University for providing expert support with native mass spectrometry experiments and V. Wysocki at Ohio State University for providing expert support with native mass spectrometry experiments. We especially thank D. Sahtoe for all the scientific support and many insightful discussions. Thanks to F. Praetorius for the brainstorming sessions dedicated to designing de novo protein motors. An additional thanks to A. Bera, M. Bick, and J. Decarreau for support in crystallography and optical microscopy respectively. Thanks to T. Daniel and L. Ceze for all the great interactions and fascinating ideas and discussions regarding the design of protein nanomachines, either computational or mechanical. Thanks to B. Coventry for very helpful advice and computational help and L. Stewart for expert help, advice, perspectives, and discussions. **Funding:** This work was support by National Science Foundation (NSF) award 1629214 (D.B.); a generous gift from the Audacious Project (D.B.); the Open Philanthropy Project Improving Protein Design Fund (D.B.); University of Washington Arnold and Mabel Beckman cryo-EM center (D.B., D.V., J.K., and J.Q.); National Institute of Allergy and Infectious Diseases grants DP1AI158186 and HHSN272201700059C (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.); a Human

Frontiers Science Program Long-Term Fellowship (A.C.); a Washington Research Foundation Senior Fellowship (A.C.); Howard Hughes Medical Institute research (A.C.); Howard Hughes Medical Institute Hanna Gray fellowship grant GT11817 (N.B.). An S10 award (S10OD032290 to D.B., D.V., J.K., and J.Q.) funded the purchase of a Glacios microscope. SAXS data were collected at the Advanced Light Source (ALS) SIBYLS beamline on behalf of US DOE-BER, through the Integrated Diffraction Analysis Technologies (IDAT) program. Additional support comes from the NIGMS project ALS-ENABLE (P30 GM124169) and a High-End Instrumentation Grant (S10OD018483). The Berkeley Center for Structural Biology is supported in part by the National Institutes of Health (NIH), National Institute of General Medical Sciences, and the Howard Hughes Medical Institute. This research used resources of the ALS, a US DOE Office of Science User Facility under contract DE-AC02-05CH11231. Some of this work was performed at the Pacific Northwest Center for Cryo-EM (PNCC), which was supported by NIH grant U24GM129547 and performed at the PNCC at Oregon Health & Science University and accessed through EMSL (grid.436923.9), a DOE Office of Science User Facility sponsored by the Office of Biological and Environmental Research. Molecular graphics and analyses were performed with UCSF Chimera, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311. **Author contributions:** Conceptualization: A.C. and D.B. Methodology: A.C., D.B., J.H., and J.K. Software: A.C., Y.H., C.X., S.E.B., G.U., U.N., P.B., D.B., D.S., A.M., N.K., W.S., and N.B. Validation: A.C., D.B., J.H., J.K., and Y.H. Formal analysis: A.C., J.H., and N.B. Investigation: A.C., J.H., N.B., Y.-J.P., A.N., D.N., and J.Q. Resources: A.C., D.B., J.H., J.K., and J.Q. Data curation: A.C., J.H., and Y.H. Writing – original draft: A.C. Writing – review & editing: A.C., D.B., J.H., Y.H., and J.K. Visualization: A.C., J.H., and N.B. Supervision: D.B. and J.K. Project administration: A.C. and D.B. Funding acquisition: D.B., J.K., D.V., A.C., and Y.H. **Competing interests:** A.C., D.B., J.H., J.K., N.B., and Y.H. are inventors on a provisional patent application submitted by the University of Washington for the design, composition, and function of the proteins created in this study. **Data and materials availability:** All data are available in the main text or the supplementary materials. All the EM maps have been deposited in the Electron Microscopy Data Bank (accession codes EMD-25575, EMD-25576, EMD-25577, EMD-25578, EMD-25579, and EMD-25580). All code is available from GitHub (https://github.com/alexiscourbet/) and is archived on Zenodo (*34*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abm1183
Materials and Methods
Figs. S1 to S22
Tables S1 to S3
References (*35*–*59*)
Movies S1 to S5
Data S1 to S5
MDAR Reproducibility Checklist

26 August 2021; accepted 21 March 2022
10.1126/science.abm1183

REPORTS

CATALYSIS

Single-step hydrogen production from NH_3, CH_4, and biogas in stacked proton ceramic reactors

Daniel Clark[1], Harald Malerød-Fjeld[1], Michael Budd[1], Irene Yuste-Tirados[1,2], Dustin Beeaff[1], Simen Aamodt[1], Kevin Nguyen[1], Luca Ansaloni[3], Thijs Peters[3], Per K. Vestre[1], Dimitrios K. Pappas[1], María I. Valls[4], Sonia Remiro-Buenamañana[4], Truls Norby[2], Tor S. Bjørheim[1], Jose M. Serra[4]*, Christian Kjølseth[1]*

Proton ceramic reactors offer efficient extraction of hydrogen from ammonia, methane, and biogas by coupling endothermic reforming reactions with heat from electrochemical gas separation and compression. Preserving this efficiency in scale-up from cell to stack level poses challenges to the distribution of heat and gas flows and electric current throughout a robust functional design. Here, we demonstrate a 36-cell well-balanced reactor stack enabled by a new interconnect that achieves complete conversion of methane with more than 99% recovery to pressurized hydrogen, leaving a concentrated stream of carbon dioxide. Comparable cell performance was also achieved with ammonia, and the operation was confirmed at pressures exceeding 140 bars. The stacking of proton ceramic reactors into practical thermo-electrochemical devices demonstrates their potential in efficient hydrogen production.

Hydrogen can be produced from CH_4-rich streams through steam reforming and water-gas shift (SMR+WGS, $CH_4 + 2H_2O = CO_2 + 4H_2$, $\Delta_{rxn}H° = 164.7$ kJ/mol) or from the emerging C-free H-carrier NH_3 through ammonia dehydrogenation (ADH, $NH_3 = 1/2N_2 + 3/2H_2$, $\Delta_{rxn}H° = 45.9$ kJ/mol) (*1*–*3*). In a conventional multistage H_2 production process, fuel combustion generates the heat for these endothermic reactions, and subsequent separation and compression occur downstream by pressure-swing adsorption and mechanical compressors. Efficiencies typically improve with scale, which favors large, centralized processes over distributed H_2 production for energy-carrier applications (*4*).

H_2 can also be separated and compressed electrochemically with proton (H^+)–conducting ceramic membranes such as Y-doped $BaZrO_3$-$BaCeO_3$ solid solutions (BZCY), which are functional and stable over a wide range of temperatures (300° to 800°C) and chemical environments (*5*–*10*). Proton ceramic electrochemical reactors (PCERs) extract pure H_2 from gas mixtures by electrolytically pumping protons across the membrane (Fig. 1A). These offer process intensification (*9*) by integrating reactions such as SMR+WGS or ADH with H_2 separation and compression, high energy efficiencies by supplying heat electrically (*3*), and reduced CO_2 emissions when that electricity is renewable (*11*).

As for any compression process, the work associated with electrochemical H_2 compression is minimized by operating isothermally (*12*). In a continuous-flow type of PCER, the compression ratio and associated entropy difference of H_2 across the membrane increase with the extent of separation along the reactor (Fig. 1B and fig. S1). This entropy difference is expelled as heat ($Q = T\Delta S$) during the compression process, which, if left unbalanced, leads to gradually increasing temperature along the reactor and in turn larger electric energy consumption per kilogram of compressed H_2 (Fig. 1B).

More efficient isothermal operation can be achieved by locally balancing this heat evolution with a reversible heat sink such as an endothermic chemical reaction (Fig. 1C) (*9*). However, matching the spatial distribution of heat from compression with the extent of chemical reactions throughout a stacked reactor poses one of the main hurdles in scaling PCERs from laboratory to commercial scale. Furthermore, scaled reactors with efficient current distribution have been hindered by the lack of interconnect materials with high electrical conductivity and chemical stability up to 800°C that match the low thermal expansion coefficient (8×10^{-6} K^{-1}) of the preferred proton conductor BZCY. Mismatches in thermal expansion of reactor components can lead to mechanical stresses and electrical contact failures during thermal cycling.

We present an optimized reactor architecture aided by multiphysics simulations and a new expansion-matched metal/glass-ceramic composite interconnect (IC) enabling deployable modular PCER stacks that retain the energy efficiencies and H_2 recoveries of single cells (*9*) while achieving a 36-fold increased H_2 production capacity. The reactor is designed with gas flows that allow internal heat exchange from exothermic to endothermic processes to minimize auxiliary heat input.

Our PCER can separate H_2 by decompression while recovering electric energy or by

[1]CoorsTek Membrane Sciences AS, 0349 Oslo, Norway. [2]Department of Chemistry, Centre for Materials Science and Nanotechnology, University of Oslo, 0316 Oslo, Norway. [3]Department of Sustainable Energy Technology, SINTEF Industry, 0314 Oslo, Norway. [4]Instituto de Tecnologia Química, Consejo Superior de Investigaciones Científicas-Universitat Politècnica de València, 46022 Valencia, Spain.
***Corresponding author. Email: jmserra@itq.upv.es (J.M.S.); ckjolseth@coorstek.com (C.K.)**

compression through supply of electric energy (Fig. 1D) at pressures up to 141 bars, illustrating the range of compression ratios and associated cell voltages that can be achieved throughout the PCER stack length. The PCER stack is a series of six barrels, each with six single cells connected electrically in parallel (Fig. 2A and fig. S4A), that uses a newly developed Ni-based glass-ceramic composite ICs (*13*). A conductive washer of the same material is placed between the end of each membrane segment and the IC plate (Fig. 2, E and F). Pure metals with higher thermal expansions, such as Ni or Cu, are prone to loss of electric contact during thermal cycling (fig. S5A, Ni washer) because of an expansion mismatch with the IC, glass-ceramic, and membrane segment (fig. S5B). Our expansion-matched Ni/glass-ceramic composite washer is applied in a partially heat-treated condition as a sintered Ni/glass composite rather than a fully heat-treated Ni/glass-ceramic composite. This means that the washer can deform under the load applied during the heating phase of the sealing cycle by virtue of viscous flow in the glassy matrix phase in the washer and maintain intimate surface contact with both components.

The glassy matrix phase wets the ceramic phases in both the tubular cell support and the IC and produces a mechanically strong bond. By the end of the sealing cycle, the glassy matrix phase in the washer crystallizes to produce a matched expansion Ni/glass-ceramic composite bridge, which retains excellent electrical continuity between the cell and the IC throughout subsequent thermal cycling. The adopted IC material exhibits conductivities >2500 S/cm at 750°C and thermal expansion coefficients in a close-to-perfect match with the thermal expansion of the membrane support (Fig. 2D and fig. S5B), ensuring efficient current distribution throughout the stack and mechanical robustness. The IC is chemically stable under reducing and CO_2-rich atmospheres, but can also be fitted with more oxidation-resistant metallic components such as Ag for operation under oxidizing conditions (fig. S5, C to E). The absence of Cr furthermore eliminates degradation issues related to formation of resistive Cr_2O_3 scales or evaporation of volatile Cr species during long-term operation at high temperatures.

During operation, individual cells will be net endothermic or exothermic depending on the degree of reaction and H_2 separation and compression throughout the stack length (Figs. 1D and 2B), necessitating internal heat exchange. To guide the optimal design of the stack, we adopted a three-dimensional multiphysics model integrating coupled gas flows, heat transfer, current distribution, and reaction kinetics for SMR+WGS and ADH that captures the behavior of the stack from single cell to stack level (*13*). Our stack is designed with a U-bend type of gas flow pattern achieved by a manifold that distributes the incoming gas to three of the six gas channels in the stack while combining the three corresponding exhaust streams. For the axial type design, the fast kinetics of SMR and ADH concentrates the heat consumed by the endothermic reactions to the initial segments of the stack, whereas the heat caused by compression primarily evolves in the latter segments. This leads to temperature increase along the reactor length (Figs. 1A and 3, B and C) which in turn increases the cell Nernst voltage and compression work (Fig. 3D). Our U-bend design, however, mitigates this mismatch by spatially balancing the heat production from compression with the heat consumption of the reactions enabling a more uniform temperature profile (Fig. 3A and fig. S6, A and B). This in turn lowers cell Nernst voltages and thus the required compression work (Fig. 3F and fig. S6C). Coupled with the high performing IC, the U-bend design allows currents (i.e., hydrogen fluxes) to self-regulate according to the local Nernst voltage (fig. S4). For anhydrous ADH, SMR+WGS and biogas this is particularly evident in the initial segment where the reaction is concentrated because of fast kinetics (figs. S7 to S9). The slower reaction kinetics of aqueous ADH on the other hand distributes the reaction over a larger portion of the stack (fig. S10).

To experimentally demonstrate integration of reactions beyond SMR+WGS in our PCER stack (*9*), single cells were operated with NH_3 in both anhydrous and aqueous form (*13*). The cells achieve >97% conversion of NH_3 even at open-circuit conditions (Fig. 4A) and near 100% conversions at high H_2 recoveries thus leaving an effluent stream virtually free of residual NH_3. The cells demonstrate comparable performance with anhydrous and aqueous NH_3, CH_4, and biogas, retaining near faradaic behavior to above 0.7 A/cm^2 (fig. S11),

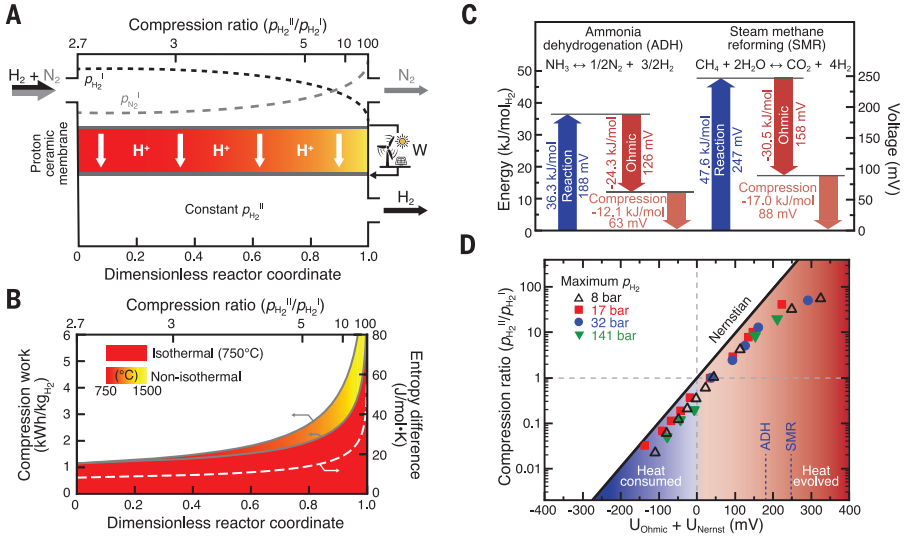


Fig. 1. H_2 separation and compression using PCERs. (**A**) Schematic of a proton ceramic electrochemical continuous-flow reactor illustrating H_2 separation from an $H_2 + N_2$ mixture. The local H_2 compression ratio (top axis) increases as $p^{I}_{H_2}$ decreases upon H_2 extraction along the reactor length (assuming a constant $p^{II}_{H_2}$), leading to a corresponding temperature increase. (**B**) Compression work for isothermal and non-isothermal H_2 separation from an $N_2 + H_2$ mixture. The local compression ratio for H_2 and the associated entropy difference $\Delta S(x) = R\ln\left(p^{II}_{H_2}/p^{I}_{H_2}(x)\right)$ increase along the reactor coordinate, which leads to an increase in compression work and heat expelled from the compression process [$w_{el}(x) = Q(x) = T\Delta S(x)$]. If left unbalanced, this heat increases the temperature throughout the reactor, particularly in the latter parts, resulting in higher compression work than would be ideal for isothermal operation. (**C**) Energy balance and correlated voltages for thermally balanced operating modes, which include reaction ($E_R = \Delta H_R$ and $U_R = \Delta H_R/nF$), charge transport ($E_C = iRnF$ and $U_{ohmic} = iR$), and compression [$E_{co} = U_{Nernst}nF$ and $U_{Nernst} = RT/nF \ln\left(p^{II}_{H_2}/p^{I}_{H_2}\right)$] for SMR+WGS at 750°C with an H_2 compression ratio of 7.4 and NH_3 cracking at 650°C with an H_2 compression ratio of 5.4. See also fig. S2 for a decompression mode of operation. (**D**) Compression ratio as a function of $U_{Ohmic} + U_{Nernst}$ at 750°C for a maximum pH_2 from 8 to 141 bars measured using representative PCER single cells (fig. S3) at i = 50 mA/cm^2, illustrating the different operation modes. The compression ratio range was covered by adjusting the minimum pH_2 as well as gas flows to ensure a low degree of H_2 extraction/dilution. The blue region consumes and the red region evolves heat.

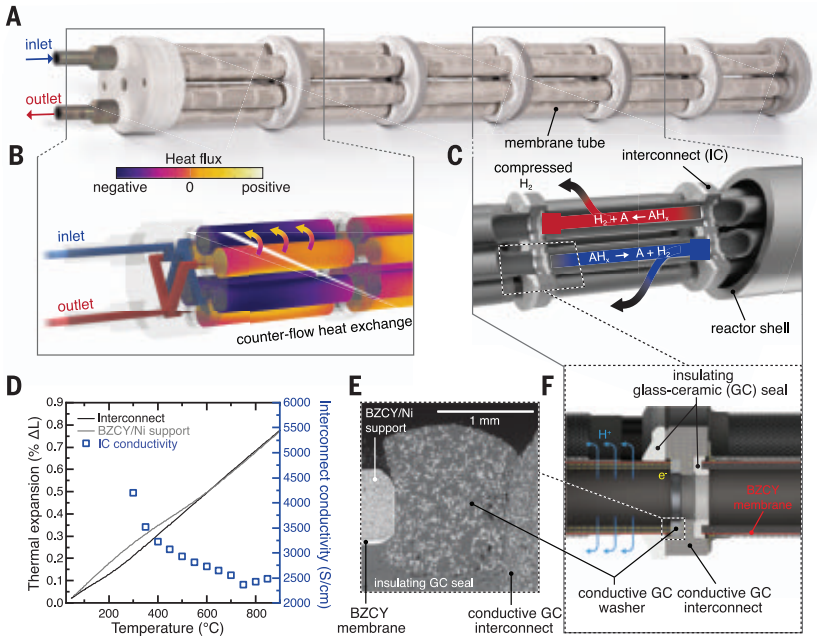
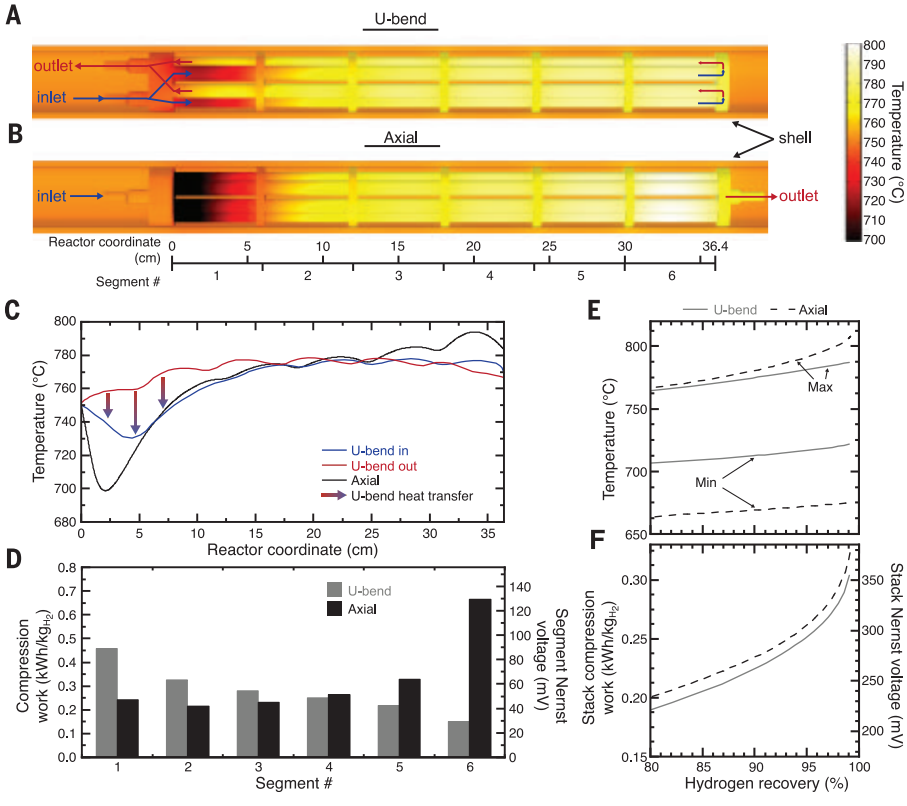


Fig. 2. PCER stack for electrochemical H_2 production. (**A**) PCER stack (dimensions: height 43 cm, diameter 4 cm). (**B**) Schematic of microthermal heat integration with outward heat flux from the cells. (**C**) Schematic of U-bend type of gas flow of the generic molecule AH_x reacting to form H_2, which is electrochemically extracted as H^+ through the membrane and recovered as compressed H_2 in the outer chamber. (**D**) Thermal expansion upon cooling of the BZCY/Ni support and the IC, and IC electrical conductivity as a function of temperature. (**E**) Scanning electron micrograph cross section of the interface between the BZCY/Ni support and IC, connected by a conductive glass-ceramic washer. (**F**) Schematics of IC and washer assembly.

reflecting the catalytic versatility of the porous Ni-BZCY support. With CH_4, the single cells even achieved >90% faradaic efficiency up to 7.4 A/cm^2 (corresponding to a H_2 flux of 47 normal milliliters-per-minute square centimeter) (fig. S12) thus doubling the H_2 production capacity to-date with these materials (*9*).

The 36-cell PCER stack achieves nearly full CH_4 conversion and high H_2 recoveries (>99%; Fig. 4 and figs. S13 and S14) for CH_4 and biogas, enabling complete equilibrium shift and a CO_2 rich effluent stream for facile carbon capture. The series and parallel design of the stack facilitates an effective aggregated current of up to 400 A (i > 0.73 A/cm^2) with a H_2 production rate up to 0.34 kg/day from CH_4, 0.31 kg/day from biogas, and 0.34 kg/day from simulated fully decomposed NH_3 streams (Fig. 4D). We furthermore demonstrate H_2 compression to 31 bar with a purity of 99.995% (Fig. 4E) facilitated for additional compression and use. The PCER stack shows promising stability, retaining a H_2 production rate of 2 normal liters per minute after 1400 hours of operation (fig. S14B). Both the H_2 production rate [0.34 versus 0.025 kg/day (*9*)] and active area [584 versus 81 cm^2 (*14*)] greatly surpass those of any reported for proton ceramic applications. Moreover, the PCER stacks have demonstrated that it is possible to deliver high-pressure H_2 at high purity and a CO_2 rich effluent at a hydrogen recovery and methane conversion >99%, which

Fig. 3. Multiphysics simulations of PCER stack thermally balanced operation. Multiphysics simulations for a stack operating at 750°C external temperature with 20 bars of total pressure on both sides of the membrane and a mean current density of 0.60 A/cm^2. Feed: 28.6% CH_4 (0.597 NL/min), 71.4% H_2O. Sweep: H_2O (0.18 g/min). (**A**) Simulated temperature fields in a U-bend PCER stack architecture, also showing the gas inlet and outlet flow distribution. (**B**) Simulated temperature fields in an axial PCER stack architecture. (**C**) Temperature profiles on the reforming side in the axial and U-bend architecture along the reactor length. The thermal balancing by heat transfer between first cell (net endothermic) and last cell (net exothermic) for U-bend PCER is illustrated by vertical arrows. (**D**) Mean compression work and Nernst voltage for each segment along the reactor length for the U-bend and axial PCERs. The values were obtained by integrating the compression work over each segment divided by the corresponding total flux or current. (**E**) Effect of H_2 recovery on the temperature distribution in U-bend and axial PCERs. (**F**) Effect of H_2 recovery on the mean stack compression work and Nernst voltage for U-bend and axial PCERs.



is highly competitive with Pd-based membrane reformers (table S2). These key performance indicators build the foundation for highly energy-efficient hydrogen production at system level.

System modeling (*13*) of a 1 ton/day distributed H_2 production plant adopting our PCER stack (figs. S15 to S21) reveals that efficiencies of 91% for CH_4 and as high as 95% for anhydrous NH_3 can be achieved by virtue of microthermal integration and downstream heat recovery. Furthermore, the PCER delivers a concentrated and pressurized stream of CO_2 when operated on methane or biogas (Fig. 4F) that can be purified and liquefied by cryogenic distillation, eliminating the need for complex downstream absorption-based CO_2 capture.

The high degree of process intensification achieved by our PCER stacks enables a fuel-flexible energy-efficient alternative to established technologies for distributed H_2 production. Using a California 2020 electric grid carbon intensity scenario (82.92 g_{CO2}/MJ_{elec}; see table S3 for references), H_2 production with PCERs using CH_4 as fuel would operate at lower emissions (75.7 vs. 124.1 g_{CO2}/MJ_{H2}) than water electrolysis powered by grid electricity, even without CO_2 sequestration. With decarbonization of the electric grid, CO_2 sequestration is required for methane reforming to remain competitive with water electrolysis. In a California 2050 grid scenario, PCERs can produce H_2 from CH_4 with lower CO_2-emissions than water electrolysis (18.7 vs. 26.2 g_{CO2}/MJ_{H2}) when CO_2 is sequestrated. PCERs operated on biogas even offer H_2 production with net-negative carbon emission, as CH_4 from a biogenic process is considered carbon-neutral. Calculated scenarios have used the CA GREET model (*15*) which includes fugitive methane emissions from natural gas production that can be important (*16*).

To illustrate the practical implications of the PCER technology, comparable well-to-wheel emissions for battery electric vehicles (BEVs), internal combustion engines (ICEs) with diesel fuel, and H_2 fuel-cell electric vehicles (FCEVs) are provided in figs. S22 and S23 with sensitivity to electric grid carbon intensity. In the California 2050 scenario, the emissions of FCEVs (14.6 g_{CO2}/km) using H_2 produced from CH_4 with PCERs including CO_2 sequestration are 90% lower than those of ICE with diesel fuel (145.4 g_{CO2}/km) and 26% lower than FCEVs using H_2 from grid-powered water electrolysis (19.8 g_{CO2}/km). NH_3-based H_2 can offer reduced emissions compared with on-site electrolysis for a wide range of electric grid carbon intensities, making FCEVs fueled with NH_3-based H_2 directly comparable to BEVs in terms of CO_2 emissions (6.3 g_{CO2}/km, a reduction of 21% compared with BEV in the California 2050 scenario). Here, NH_3 is assumed produced at off-site locations with favorable renewable energy resources and transported as a liquid to the fueling station where efficient ADH and separation to H_2 takes place using the PCER technology.

The growth of a new energy technology can be limited by access to raw materials. A detailed examination of raw materials' usage of the PCER stack (fig. S24) shows it is composed of nonprecious, earth-abundant materials, suggesting no material availability setbacks for scaling.

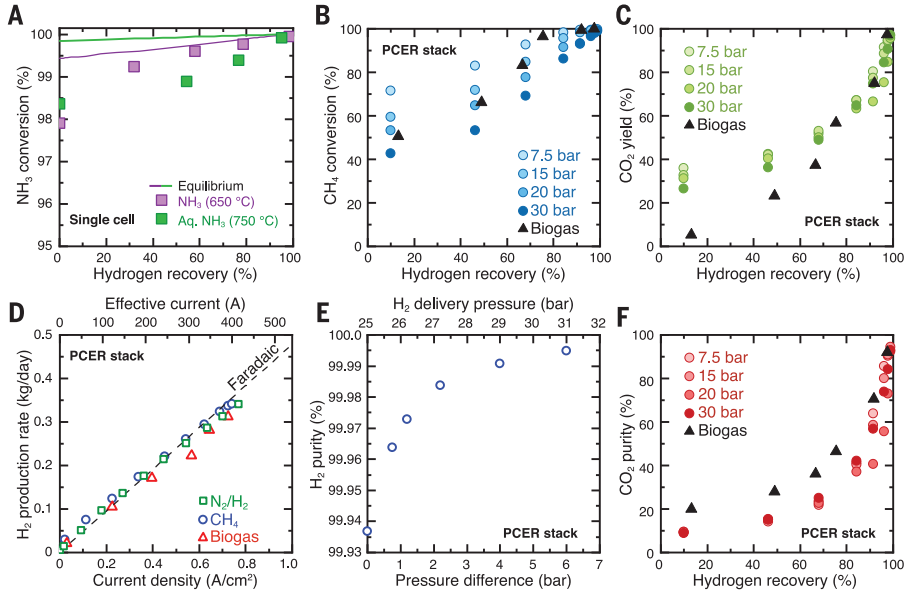


Fig. 4. PCER single-cell and stack performance. (**A**) NH_3 conversion as a function of H_2 recovery, measured on a representative single cell (fig. S3) at 650°C and 10 bars (pNH_3 = 7.25 bars; pH_2O = 2.75 bars), and aqueous NH_3 at 750°C and 10 bars (pNH_3 = 3.1 bars; pH_2O = 5.8 bars; and p_{inert} = 1.1 bars). Purple and green lines show the equilibrium conversion for NH_3 and aqueous NH_3, respectively. (**B**) and (**C**) CH_4 conversion and yield of CO_2 versus H_2 recovery, respectively, of PCER stack at 750°C. (**D**) H_2 production rate as a function of applied current density for the stack with N_2/H_2 mixture simulating complete NH_3 decomposition (750°C, 10 bars), methane (800°C, 15 bars, S/C = 2.5), and biogas (750°C, 20 bars, S/C = 2.5). Effective current is calculated as current density × PCER stack area (36 × 15 cm^2) and applied current as effective current/6 due to the series and parallel electric architecture. (**E** and **F**) H_2 purity (dry basis) versus H_2 delivery pressure and differential pressure across the membranes (E) and CO_2 purity versus H_2 recovery (F) for SMR+WGS in the stack at 750°C. Reforming side pressure = 25 bars, H_2 side pressure = 25 to 31 bars. Current density = 0.69 A/cm^2.

REFERENCES AND NOTES

1. R. F. Service, *Science* **345**, 610–610 (2014).
2. I. Staffell *et al.*, *Energy Environ. Sci.* **12**, 463–491 (2019).
3. S. T. Wismann *et al.*, *Science* **364**, 756–759 (2019).
4. G. Grashoff, C. Pilkington, C. Corti, *Platin. Met. Rev.* **27**, 157–169 (1983).
5. J. M. Serra, *Nat. Energy* **4**, 178–179 (2019).
6. S. Choi, T. C. Davenport, S. M. Haile, *Energy Environ. Sci.* **12**, 206–215 (2019).
7. C. Duan *et al.*, *Nature* **557**, 217–222 (2018).
8. C. Duan *et al.*, *Science* **349**, 1321–1326 (2015).
9. H. Malerød-Fjeld *et al.*, *Nat. Energy* **2**, 923–931 (2017).
10. H. An *et al.*, *Nat. Energy* **3**, 870–875 (2018).
11. J. L. Barton, *Science* **368**, 1181–1182 (2020).
12. B. L. Kee *et al.*, *Membranes (Basel)* **9**, 77 (2019).
13. Materials and methods are available as supplementary materials.
14. R. J. Braun *et al.*, *ECS Trans.* **91**, 997–1008 (2019).
15. California Air Resources Board, "CA-GREET3.0 lookup table pathways: Technical support documentation" (California Air Resources Board, 2018); https://ww2.arb.ca.gov/sites/default/files/classic//fuels/lcfs/ca-greet/lut-doc.pdf.
16. R. W. Howarth, M. Z. Jacobson, *Energy Sci. Eng.* **9**, 1676–1687 (2021).

ACKNOWLEDGMENTS

Funding: This work was supported by Norway's Ministry of Petroleum and Energy through the Gassnova project CLIMIT grant 618191 in partnership with Engie SA, Equinor, ExxonMobil, Saudi Aramco, Shell, and TotalEnergies and the Research Council of Norway NANO2021 project DynaPro grant 296548. **Author contributions:** Conceptualization: D.C., H.M.-F., T.P., P.K.V., T.S.B., J.M.S., C.K.; Investigation: D.C., H.M.-F., M.B., I.Y.-T., D.B., K.N., L.A., T.P., D.K.P., M.I.V., S.R.-B., T.S.B., C.K.; Methodology: D.C., H.M.-F., M.B., I.Y.-T., S.A., L.A., T.P., P.K.V., D.K.P., S.R.-B., C.K.; Resources: M.B., D.B., K.N., M.I.V.; Software: I.Y.-T., S.A.; Supervision: T.N., T.P., J.M.S., C.K.; Writing – original draft: D.C., H.M.-F., M.B., I.Y.-T., S.A., P.K.V., D.K.P., T.N., T.S.B., J.M.S., C.K.; Writing – review and editing: D.C., H.M.-F., M.B., I.Y.-T., T.S.B., D.B., S.A., L.A., T.P., P.K.V., D.K.P., S.R.-B., T.N., T.S.B., J.M.S., C.K. **Competing interests:** D.C., H.M.-F., M.B., I.Y.-T., D.B., S.A., K.N., D.K.P., T.S.B., and C.K. are employed by CoorsTek Membrane Sciences (CTMS). CTMS has filed relevant patent application PCT/EP2017/076340. T.N. is a member of the CTMS board. I.Y.-T.'s doctoral studies at the University of Oslo (UiO) are partially funded by CTMS. The remaining authors declare no competing interests. **Data and materials availability:** All data are available in the main text or the supplementary materials. Experimental data are available online at http://hdl.handle.net/10251/181917.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abj3951
Materials and Methods
Figs. S1 to S30
Tables S1 to S7
References (*17*–*50*)

16 June 2021; resubmitted 28 October 2021
Accepted 10 February 2022
10.1126/science.abj3951

CELL BIOLOGY

Cell division in tissues enables macrophage infiltration

Maria Akhmanova[1]*, Shamsi Emtenani[1]†, Daniel Krueger[2]‡§, Attila Gyoergy[1]‡, Mariana Guarda[1], Mikhail Vlasov[3], Fedor Vlasov[3], Andrei Akopian[3], Aparna Ratheesh[1]¶, Stefano De Renzis[2], Daria E. Siekhaus[1]*

Cells migrate through crowded microenvironments within tissues during normal development, immune response, and cancer metastasis. Although migration through pores and tracks in the extracellular matrix (ECM) has been well studied, little is known about cellular traversal into confining cell-dense tissues. We find that embryonic tissue invasion by *Drosophila* macrophages requires division of an epithelial ectodermal cell at the site of entry. Dividing ectodermal cells disassemble ECM attachment formed by integrin-mediated focal adhesions next to mesodermal cells, allowing macrophages to move their nuclei ahead and invade between two immediately adjacent tissues. Invasion efficiency depends on division frequency, but reduction of adhesion strength allows macrophage entry independently of division. This work demonstrates that tissue dynamics can regulate cellular infiltration.

Cell dissemination into tissues is fundamental for the formation and maintenance of complex organisms (*1*–*3*). During vertebrate development, neural crest cells move into tissues to form different tissue types (*3*); immune cells enter organs to regulate tissue function and combat infection (*4*); cancer cells traverse into other organs during metastasis (*2*). These various embryonic and adult environments contain closely packed cells adherent to one another or to the extracellular matrix (ECM) that lies between them (*1*, *5*). Despite its importance, invasion into such cell-dense tissues is poorly understood.

Macrophages are scavenger cells that invade tissues early on in development to establish residency and patrol organs (*6*). In the early *Drosophila* embryo, macrophages follow guidance cues (*7*) and invade the germ band (GB) at an entry point (EP) in the acute angle between the basal side of the ectoderm and the mesoderm surface (Fig. 1A and fig. S1A). They progress farther between the ectoderm and the mesoderm, which are juxtaposed across a thin layer of ECM (*8*). During this invasion, macrophages move as a chain, maintaining the separation of the two tissues established by the pioneer (Fig. 1A′; fig. S1, A′ to D; and movie S1). This first macrophage requires ~20 min to enter through the tissue barrier, using αPS2- and βPS-integrins that bind to laminin (*9*, *10*). Matrix metalloproteolytic ECM degradation does not affect the efficiency of entry (*10*) but macrophage-specific programs do (*11*–*13*). How the dynamics and properties of surrounding cells influence macrophage tissue invasion remains unclear.

We tracked macrophage nuclei (Fig. 1B and fig. S1E), the stiffest of the organelles. We plotted the velocity of the first macrophage's nucleus relative to the entry point over time and defined the moment of entry between the two tissues by its velocity jump when close to the entry point (fig. S1, F and G). As the first macrophage entered, one or two ectodermal cells adjacent to the entry point had become round, a shape change that starts at the beginning of mitosis (*14*), or had progressed to become two connected smaller rounded daughters, formed by the completion of division (Fig. 1, C and D; fig. S1, H and I; and movie S2). By analyzing the division profiles of 20 time-lapse movies, we observed that macrophage entry always occurred during the division (fig. S2, A and C) or rounding (fig. S2, B and D) of a flanking ectodermal cell (fig. S2E). We quantified how much time the ectodermal cells at the entry point spent in three categories: (i) at least one of the two cells is dividing; (ii) at least one cell is rounding; or (iii) the interdivision phase, during which none of these cells are mitotic (fig. S2F). If macrophage invasion occurred randomly, macrophages would enter during the interdivision phase in more than half of the embryos; however, we never observed entry at this time (Fig. 1E). Within an embryo (when viewed in a dorsal orientation), there are two bilaterally symmetrical entry points next to the hindgut. Macrophages arrived at these two locations almost simultaneously; however, they did not enter at the same time if ectodermal divisions at both points occurred asynchronously (fig. S3). We also found no divisions of mesodermal cells at the entry site during invasion (fig. S4), perhaps because the mesodermal division frequency is much lower than in the ectoderm (fig. S5). Hence, in wild-type embryos, macrophages enter the GB tissue only when an adjacent ectodermal cell is dividing, suggesting that this is a permissive event.

To investigate the ectoderm–mesoderm interface structure, we stained embryos for the ECM component laminin, secreted there previously by the mesoderm and also expressed within entering macrophages (*15*). Quantification shows that more laminin remains

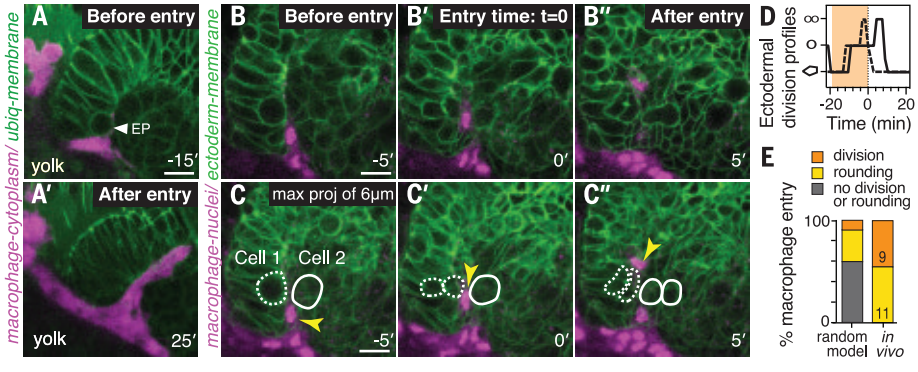


Fig. 1. Macrophage invasion correlates with ectodermal cell division at the entry point. (**A** and **A′**) Still images of cytoplasm-labeled macrophages at the entry point (EP, white triangle) and after invading as a chain. ubiq-membrane, ubiquitous membrane marker. (**B** and **C**) Single plane (B) still images and (C) corresponding maximum projections showing a nuclear-labeled macrophage (yellow arrowhead) at the EP with ectodermal cell 1 (dashed outline) adjacent to, and cell 2 (solid outline) above, this entering macrophage. The outlined ectodermal cells mitotically round and divide. Scale bars, 10 μm. (**D**) Cell 1 and 2 division profiles in time. Peak indicates division; middle level, rounded cell; and lower level, polygonal shape. Time of macrophage entry, $t = 0$. Shading indicates macrophage at EP. (**E**) Quantification of macrophages entering together with ectodermal cell cycle phases. (Left) Theoretical estimate assuming random entry proportional to average duration of each phase, presented in fig. S2F. (Right) In vivo determined entry for 20 embryos (fig. S2, A and B).

[1]Institute of Science and Technology Austria (IST Austria), Klosterneuburg, Austria. [2]Developmental Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. [3]Bundesgymnasium Klosterneuburg, Klosterneuburg, Austria.
*Corresponding author. Email: maria.akhmanova@gmail.com (M.A.); daria.siekhaus@ist.ac.at (D.E.S.)
†Present address: Institute of Molecular Biotechnology, Austrian Academy of Sciences, Vienna Biocenter, Vienna, Austria.
‡These authors contributed equally to this work.
§Present address: Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences, Utrecht, Netherlands.
¶Present address: Centre for Mechanochemical Cell Biology, Warwick Medical School, University of Warwick, Coventry, UK.

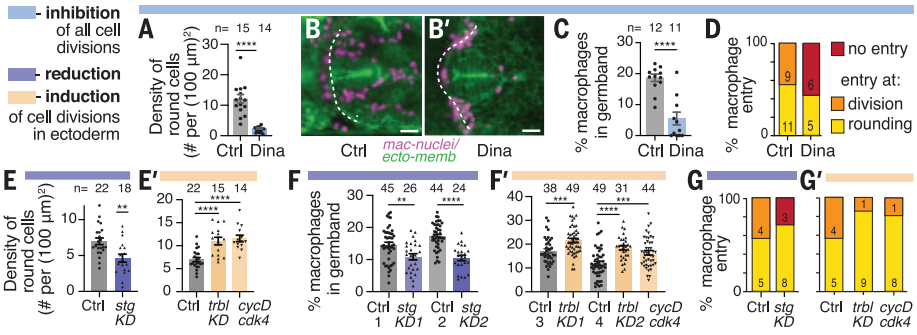


Fig. 2. Macrophage entry requires ectodermal mitotic rounding. (**A**) Rounded cell density in the GB tissue at macrophage entry time and (**B** and **C**) percent of all macrophages that were found inside the GB as GB retraction started in control dimethyl sulfoxide–injected embryos (Ctrl) and dinaciclib-injected embryos (Dina). (B) and (B′) show dorsal view of maximum intensity projections of stage 12 embryos. Dashed line indicates the GB edge. Scale bars, 20 µm. mac-nuclei, macrophage nuclei; ecto-memb, ectoderm membrane. ****P < 0.0001 [(A) and (C)]. (**D**) Quantification of macrophage entry timing. (**E** and **E′**) Rounded cell density in the GB altered by cell cycle regulators. **P = 0.0021; ****P < 0.0001. (**F** and **F′**) Percent of all macrophages that were found inside the GB. **P = 0.0067, ****P < 0.0001 (F); (left to right) ***P = 0.0007, ****P < 0.0001, ***P = 0.0009 (F′). (**G** and **G′**) Quantification of macrophage entry timing. Mean ± SEM. Ten planes per embryo analyzed in (A), 5 to 10 in (E) and (E′). Mann-Whitney tests [(A) and (C)], unpaired two-tailed t tests [(E) to (F′)].

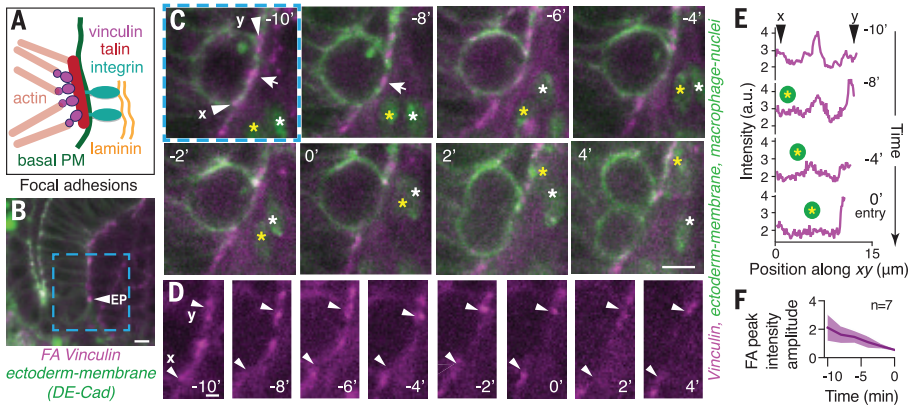


Fig. 3. Macrophages enter when FAs disassemble during ectodermal mitotic rounding. (**A**) Scheme of ectodermal FAs. (**B**) FAs in the ectoderm–mesoderm interface visualized by vinculin::mCherry (magenta) with ectodermal cells' membranes (green). DE-Cad, *Drosophila* E-cadherin. (**C**) Time-lapse imaging of ectodermal FAs during mitotic rounding and subsequent macrophage entry at t = 0 (yellow and white stars). White arrows indicate large FA that disassembles. (**D**) Vinculin::mCherry on the basal side of mitotic cell, cutout from (C). (**E**) Vinculin::mCherry intensity along the basal side of rounding cell in (C) and (D) from point x to point y (white arrowheads) over time. Green circle indicates first macrophage nucleus. a.u., arbitrary units. (**F**) Amplitude of FA peak over time. Mean ± SD (shading). Scale bars, 5 µm [(B) and (C)] and 2 µm (D).

on the mesodermal side of the first macrophage's protrusion (fig. S6). Thus, macrophages appear to enter next to the ectodermal cells and are separated from the mesoderm by the ECM.

Next, we examined whether ectodermal division is required for macrophage entry. We injected embryos with the drug dinaciclib, which inhibits cyclin-dependent kinase 1 (Cdk1) and other cyclin-dependent kinases and effectively stops the cell cycle (fig. S7, A and B) without affecting morphogenetic tissue movements (fig. S7C) or total macrophage numbers (fig. S7D). The density of rounded ectodermal cells was 83% lower after dinaciclib injection (Fig. 2A). Macrophages remained motile and moved directionally (fig. S7, E and F) toward the entry point; however, they almost completely failed to invade (Fig. 2, B and B′, and fig. S7, G and G′). We quantified macrophages inside the GB ~60 min after macrophages arrived at the entry point and normalized for the total macrophage number. In control embryos, ~20% of total macrophages had moved inside the GB by this time. However, in dinaciclib-treated embryos, only a few macrophages entered the GB (Fig. 2C). Live imaging revealed that when a few macrophages entered, they moved in at the usual site next to a round ectodermal cell that did not progress to complete division for a long period of time (fig. S7, H to K, and movie S3). Thus, pharmacological inhibition of cell division resulted in macrophages not invading at all, or entering next to the remaining mitotically rounded ectodermal cells (Fig. 2D).

To confirm these results, we inhibited division only locally, by expressing RNA interference (RNAi) constructs in the ectoderm against mRNA encoding a positive regulator of mitosis (fig. S7A), Cdc25 (Stg). This knockdown (KD) reduced the ectodermal frequency of rounded cells by ~30% (Fig. 2E). Between 25 and 40% fewer macrophages penetrated the GB upon two different *stg* KDs compared with the control (Fig. 2F and fig. S8, A and A′). In embryos expressing these RNAis, in which no divisions occurred in the GB edge for a long time, no macrophages entered (fig. S8B and movie S4, left). When we detected entry in other embryos, it was always adjacent to a remaining mitotic ectodermal cell at the normal location (fig. S8C and movie S4, right). Similar results were observed upon ectodermal overexpression of p53 (fig. S8, D and E). Therefore, reducing the frequency of ectodermal cell division decreases macrophage entry, while maintaining the entry-mitotic rounding correlation (Fig. 2G and fig. S8F).

To test whether increasing the rate of mitosis in the ectoderm would cause the opposite result, we knocked down a negative regulator of mitosis, Tribbles (Trbl), or overexpressed the G1 progression regulators Cyclin D (CycD) and Cdk4 in the ectoderm (fig. S7A). These treatments increased the density of ectodermal rounded cells before macrophage entry by 60% (Fig. 2E′) and led to a 25 to 50% increase in macrophages in the GB compared with the control (Fig. 2F′ and fig. S8, G and G′). Macrophages always entered at their usual position and next to a mitotic ectodermal cell (Fig. 2G′, fig. S8H, and movie S5). Total macrophage numbers were unchanged compared with controls (fig. S8, I and J). These results strongly suggest that the timing of ectodermal divisions is a rate-limiting factor for macrophage invasion.

To examine whether macrophages stimulate division to facilitate invasion, we expressed Hid in macrophages, inducing their apoptosis before reaching the GB (fig. S9A). However, the density of dividing cells in the GB was unaltered in these embryos, and the timing and speed of GB retraction remained unchanged (fig. S9, B and C, and movie S6).

We next asked how ectodermal mitotic rounding facilitates macrophage entry. First, we tested whether the increased cortical stiffness that causes rounding, or the free space that results from it, might help macrophages to advance (*14*, *16*). We temporarily increased cortex contractility through optogenetic recruitment of the Rho1 exchange factor RhoGEF2 to the plasma membrane (fig. S10A) (*17*), inducing rounding of ectodermal cells in a small region (50 μm by 40 μm by 20 μm) at the entry site ~40 min before normal macrophage entry (fig. S10, B to D) without causing de-adhesion of these cells (fig. S10, E to H). This approach, distinct from the long-lasting widespread increase in ectodermal tension that impedes macrophage invasion (*8*), caused no change in the timing of macrophage entry (fig. S10, I to K). Next, using fluorescent dextran, we examined the intercellular space that mitotic rounding opens at the entry point in wild-type embryos. This distance increased by only 30%, producing a gap narrower than the diameter of the macrophage nucleus when squeezed between tissues and much narrower than the space observed between the tissues during invasion (fig. S11). These results suggest that ectodermal cell rounding alone is not sufficient to promote macrophage entry.

The ectoderm faces the mesoderm with its basal side, forming focal adhesions (FAs) that bind through αPS1- and βPS-integrins to laminin, which could thus impede macrophage entry (Fig. 3, A and B) (*8*). FAs have been observed to disassemble during vertebrate mitotic rounding in vitro (*18*). We examined their in vivo temporal and spatial dynamics in the GB. We first visualized these FAs by staining against β-integrin, which localized to dot-like adhesive structures at the ectodermal–mesodermal interface (fig. S12, A to A″). Focal adhesion components vinculin and talin (*19*) also showed a dotted pattern along this interface (fig. S12, B to D). Once an ectodermal cell started to round, these FAs (visualized live by vinculin::mCherry) gradually disappeared, leaving only one peak remaining in the middle of the basal side (fig. S12E and Fig. 3, C and D), which gradually flattened (Fig. 3, C to E). FA disassembly did not depend on the presence of macrophages and took place in every basally dividing cell in the ectoderm (fig. S13, A to E, and movie S7); FAs reassembled after daughter cells acquired a polygonal shape (fig. S13, F and G, and movie S8). Imaging ectodermal mitotic cells at the entry point revealed that macrophages always entered after disassembly of the last adhesion spot (Fig. 3, C to F; movie S9; and fig. S14, A to F). The macrophage nucleus moved farther into the tissue between other FAs, which do not disassemble (fig. S14, B′ to F′ and G to H′). In cases where entry occurred next to the rounded daughter cells produced by division, mitosis was completed faster than macrophage nuclear translocation; however, new FAs were not established before entry (fig. S13, F and G). We conclude that FA disassembly at the entry point correlates with the macrophage nucleus's penetration between tissues.

Finally, we reduced FA components β-integrin, talin, or vinculin in the ectoderm by RNAi knockdown (fig. S15, A to F) without affecting development (fig. S15, G and H). This resulted in higher macrophage numbers in the GB (Fig. 4A and fig. S16, A to C) that entered earlier but without a change in speed compared with the control (fig. S16, D to G). In contrast, β-integrin knockdown in the mesoderm had no such effect (fig. S16H), confirming that mesoderm adhesiveness does not hinder invasion. The first macrophage could enter without ectodermal rounding or division in β-integrin or vinculin knockdown embryos, unlike in the wild type (Fig. 4B, fig. S16I, and movie S10). Even in dinaciclib-injected embryos, ectodermal β-integrin knockdown resulted in macrophages entering in the absence of rounding or dividing cells at the entry point, further suggesting that rounding itself is not necessary for invasion (fig. S16J). Thus, although we cannot exclude a contribution from other effects, we reveal the disassembly during mitosis of the focal adhesions that attach a cell to its environment as the main mechanism by which division opens the door for macrophage infiltration (fig. S17).

We have pinpointed surrounding tissue division as the crucial variable affecting the rate of entry of the first *Drosophila* macrophage. Vertebrate mitosis also leads to diminishment of focal adhesions (*18*). Our findings suggest that regulation of division during development, inflammation, or tumor growth could affect the number and placement of immune cells in tissues in a wide range of normal and disease contexts.

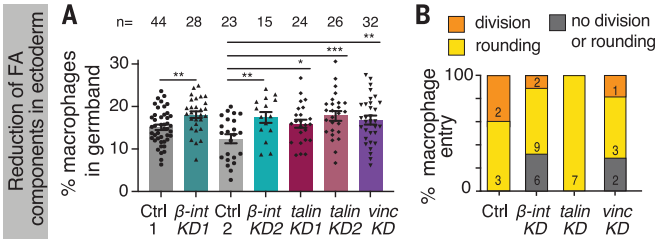


Fig. 4. Reducing ectodermal FA components enables macrophage entry without mitotic rounding. (**A**) Percent of all macrophages that were found inside the GB in the indicated knockdown embryos. Mean ± SEM. (Left to right) $^{**}P = 0.0038$, $^{**}P = 0.0052$, $^{*}P = 0.018$, $^{***}P = 0.0005$, $^{**}P = 0.0041$, unpaired *t* tests. (**B**) Quantification of macrophage entry timing.

REFERENCES AND NOTES

1. P. Friedl, B. Weigelin, *Nat. Immunol.* **9**, 960–969 (2008).
2. P. Friedl, S. Alexander, *Cell* **147**, 992–1009 (2011).
3. P. M. Kulesa, L. S. Gammill, *Dev. Biol.* **344**, 566–568 (2010).
4. K. Kierdorf, M. Prinz, F. Geissmann, E. Gomez Perdiguero, *Semin. Immunol.* **27**, 369–378 (2015).
5. T. A. Stewart, K. Hughes, D. A. Hume, F. M. Davis, *Front. Cell Dev. Biol.* **7**, 250 (2019).
6. S. Epelman, K. J. Lavine, G. J. Randolph, *Immunity* **41**, 21–35 (2014).
7. A. Ratheesh, V. Belyaeva, D. E. Siekhaus, *Curr. Opin. Cell Biol.* **36**, 71–79 (2015).
8. A. Ratheesh *et al.*, *Dev. Cell* **45**, 331–346.e7 (2018).
9. K. Comber *et al.*, *J. Cell Sci.* **126**, 3475–3484 (2013).
10. D. Siekhaus, M. Haesemeyer, O. Moffitt, R. Lehmann, *Nat. Cell Biol.* **12**, 605–610 (2010).
11. K. Valoskova *et al.*, *eLife* **8**, e41801 (2019).
12. V. Belyaeva *et al.*, *PLOS Biol.* **20**, e3001494 (2022).
13. S. Emtenani *et al.*, *EMBO J.* 10.15252/embj.2021109049 (2022).
14. A. V. Taubenberger, B. Baum, H. K. Matthews, *Front. Cell Dev. Biol.* **8**, 687 (2020).
15. Y. Matsubayashi *et al.*, *Curr. Biol.* **27**, 3526–3534.e4 (2017).
16. W. Dai *et al.*, *Science* **370**, 987–990 (2020).
17. E. Izquierdo, T. Quinkler, S. De Renzis, *Nat. Commun.* **9**, 2366 (2018).
18. C. L. Dix *et al.*, *Dev. Cell* **45**, 132–145.e3 (2018).
19. B. Klapholz, N. H. Brown, *J. Cell Sci.* **130**, 2435–2446 (2017).

ACKNOWLEDGMENTS

We thank C. Guet, T. Hurd, J. Friml, M. Fendrych, and laboratory members for comments on the manuscript. We also thank the Bioimaging Facility of IST Austria for excellent support and T. Lecuit, E. Hafen, R. Levayer, and A. Martin for fly strains. All the confocal images and movies obtained during this study are archived on a server at IST Austria and can be obtained upon request. Stocks used in this study were obtained from the Bloomington *Drosophila* Stock Center (NIH P40OD018537). **Funding:** M.A. was supported by an Austrian Science Fund FWF: Lise Meitner Fellowship M2379-B28 to M.A. and D.E.S. S.E., M.G., and A.G. were supported by internal funding from IST Austria to D.E.S. D.K. was supported by internal funding from EMBL to S.D.R. **Author contributions:** M.A. and D.E.S. conceived of the study, designed the experiments, and interpreted the results. M.A. discovered the phenomena with valuable input from A.R. M.A., S.E., A.G., and M.G. performed fly crossing, fixation, and imaging and analyzed the experiments. A.G. generated the *srpHemo-mCerulean::H2B::Dendra* line. M.A. performed live imaging and drug injection experiments. M.V., F.V., and A.A. performed fly work and crossing. A.A. wrote the data analysis script. A.R. provided PH3 staining images. D.K. and S.D.R. provided optogenetic setup and flies, and guided M.A. during optogenetic experiments. M.A., S.D.R., and D.E.S. acquired funding. M.A. and D.K. made illustrations. M.A. and D.E.S. wrote the manuscript, with contributions from S.D.R., D.K., S.E., M.G., and A.R. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data are available in the main text or the supplementary materials. Image analysis scripts are publicly available at https://github.com/Axmasha/Image_analysis_scripts and https://github.com/andrei-akopian/Data_alignment_script. All materials are available upon request.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abj0425
Materials and Methods
Figs. S1 to S17
Exact genotypes of *Drosophila* lines used in the figures
References (*20*–*41*)
MDAR Reproducibility Checklist
Movies S1 to S10
Data S1

16 April 2021; accepted 25 March 2022
10.1126/science.abj0425

SUPERCONDUCTIVITY

Superconducting spin smecticity evidencing the Fulde-Ferrell-Larkin-Ovchinnikov state in Sr_2RuO_4

K. Kinjo[1]*, M. Manago[1]†, S. Kitagawa, Z. Q. Mao[2], S. Yonezawa[1], Y. Maeno[1]‡, K. Ishida[1]*

Translational symmetry breaking is antagonistic to static fluidity but can be realized in superconductors, which host a quantum-mechanical coherent fluid formed by electron pairs. A peculiar example of such a state is the Fulde-Ferrell-Larkin-Ovchinnikov (FFLO) state, induced by a time-reversal symmetry–breaking magnetic field applied to spin-singlet superconductors. This state is intrinsically accompanied by the superconducting spin smecticity, spin density–modulated fluidity with spontaneous translational-symmetry breaking. Detection of such spin smecticity provides unambiguous evidence for the FFLO state, but its observation has been challenging. Here, we report the characteristic "double-horn" nuclear magnetic resonance spectrum in the layered superconductor Sr_2RuO_4 near its upper critical field, indicating the spatial sinusoidal modulation of spin density that is consistent with superconducting spin smecticity. Our work reveals that Sr_2RuO_4 provides a versatile platform for studying FFLO physics.

Conventional superconductivity, regarded as a quantum-mechanical coherent fluid formed by electron pairs, is spatially homogeneous in ideal situations and retains translational symmetry. However, such a homogeneous fluid state may become unstable, and the translational symmetry can be broken upon the application of time-reversal symmetry–breaking magnetic fields. One trivial example is quantized vortex lattices formed in any type II superconductors. A more exotic state may emerge under sufficient spin polarization: With spin imbalance, electron pairs with a finite total momentum are formed, resulting in spatial oscillation of the superconducting (SC) order parameter. This state is known as the Fulde-Ferrell-Larkin-Ovchinnikov (FFLO) state (*1*, *2*); the concept has been extended to other systems such as cold atoms (*3*) and neutron stars (*4*). In the FFLO state of a spin-singlet superconductor, the modulated SC order parameter is intrinsically accompanied by both fluidity and spatial spin modulation—namely "spin smecticity," which is analogous to liquid-crystal smectic phases. Moreover, this spin smecticity causes nontrivial local spin-density enhancement: Spin density near the nodes of the order parameter oscillation becomes even higher than that of the normal state (Fig. 1). An observation of this characteristic spin smecticity would provide indisputable evidence of the FFLO state.

Demonstrating unambiguously the existence of the FFLO state has proven difficult. This is because superconductivity is in most cases destroyed by vortex kinetic energy (orbital limiting) before sufficient spin polarization is induced by the Zeeman effect (Pauli limiting). A first-order transition (FOT) between the normal and SC states is a hallmark of sizable spin polarization. Moreover, FFLO states are quite fragile against impurities; only a few clean superconductors with large effective electron mass (*5*), such as quasi–two-dimensional (2D) organics (*6*) and heavy-fermion $CeCoIn_5$ (*7*), exhibit the FOT and subphases within the SC state. For $CeCoIn_5$, the observed high-field subphase (Q-phase) (*8*) is not a simple FFLO phase because it is intertwined with a magnetic order. In (*6*–*8*), bulk anomalies related to the FFLO state have been detected, but evidence more closely linked to the FFLO modulation, such as the SC spin smecticity, has not been reported. Nuclear magnetic resonance (NMR) spectrum is a promising probe with which to observe spatial spin modulation because NMR spectrum can detect the local spin susceptibility with high sensitivity (*9*–*12*).

In this work, we focused on the layered perovskite Sr_2RuO_4, which is an extremely clean and archetypal unconventional superconductor with the SC critical temperature (T_c) of 1.5 K (*13*–*15*). Its normal state is well understood as a strongly correlated Fermi-liquid state. However, the nature of the SC pairing state has not been elucidated. Recently, a decrease in the in-plane spin susceptibility in the SC state was observed (*16*) and later confirmed by other groups (*17*, *18*), ruling out the chiral spin-triplet scenario. The unchanged spin susceptibility reported by previous NMR measurements (*19*) is ascribed to the instantaneous destruction of superconductivity caused by overheating by NMR pulses. The full elucidation of the SC state of Sr_2RuO_4 is entering a new stage (*20*–*25*). One of the characteristic features observed in Sr_2RuO_4 is the FOT between the SC and normal states in the in-plane field (*26*). In addition, Sr_2RuO_4 has an exceedingly long mean-free-path l of ~3 μm for the in-plane direction, which is much longer than the in-plane SC coherence length of 0.066 μm (*14*, *15*). These features favor the emergence of the FFLO state.

Moreover, Sr_2RuO_4 has a crystallographic advantage for detecting the SC spin smecticity. There are two inequivalent oxygen sites in Sr_2RuO_4: the planar-oxygen [O(1)] and the apical-oxygen [O(2)] sites (Fig. 2A). The smaller hyperfine coupling and higher local symmetry at the O(2) site make the local NMR spectrum much sharper and stronger than that at the O(1) site. Thus, the NMR spectrum at the O(2) site provides an ideal probe for the spatial variation of spin density. In other FFLO candidates, the anomalies in the spectral width and the enhancement of the nuclear spin-lattice relaxation rate $1/T_1$ suggesting the local enhancement in the density of states were reported in the FFLO state (*9*–*11*). Such an enhancement of $1/T_1$ is considered to be a decisive dynamical signature of the FFLO state originating from the spin-polarized Andreev bound state (*9*–*12*). Nevertheless, previous attempts to probe the static FFLO spin modulations by the spectral peak splitting have not yielded a clear and unambiguous observation, mainly owing to unfavorable crystal structure and complications caused by magnetic ordering.

We used ^{17}O-NMR measurements with a very precise field-direction control to observe that the spin density ascribed to the SC pairs becomes spatially inhomogeneous, featuring spin-"dense" and spin-"thin" portions. Details of the experimental methods are described in (*27*). This spin modulation emerges in a limited field region near SC upper critical field H_{c2} and only when magnetic field H is parallel to the *ab* plane. Our observations indicate the existence of SC spin smecticity, providing unambiguous evidence for the FFLO state.

The in-plane field variation of the ^{17}O-NMR spectrum is shown in Fig. 2B at the O(2) site near H_{c2}, measured at 70 mK. All spectra were obtained by using the free-induction-decay method with the radio frequency (RF)–pulse energy small enough to avoid overheating (figs. S2 and S3). Together with time-resolved ac susceptibility measurements by using the identical sample and setup, the data confirm that the sample is globally in the SC state below H_{c2} (fig. S2) (*16*, *17*). In the following, NMR spectra are presented as a function of the Knight shift K, which is defined as $K = (f - f_0)/f_0$, where f_0 is the bare NMR frequency evaluated by $f_0 = \gamma H/2\pi$, with the nuclear gyromagnetic ratio $\gamma/2\pi$ = 5.7718 MHz/T of ^{17}O nuclei. K at the peak (K^{peak}) is proportional to the electronic spin susceptibility at the nuclear site. The NMR spectra above $\mu_0 H_{c2}$ = 1.4 T, where the sample is in the normal state, are almost symmetric. The spectra below 1.1 T

[1]Department of Physics, Graduate School of Science, Kyoto University, Kyoto, Japan. [2]Department of Physics, Pennsylvania State University, State College, PA, USA.
*Corresponding author. Email: kinjo.katsuki.63v@st.kyoto-u.ac.jp (K.K.); kishida@scphys.kyoto-u.ac.jp (K.I.)
†Present address: Department of Physics and Material Science, Shimane University, Shimane, Japan.
‡Present address: Toyota Riken–Kyoto University Research Center (TRiKUC), Kyoto, Japan.

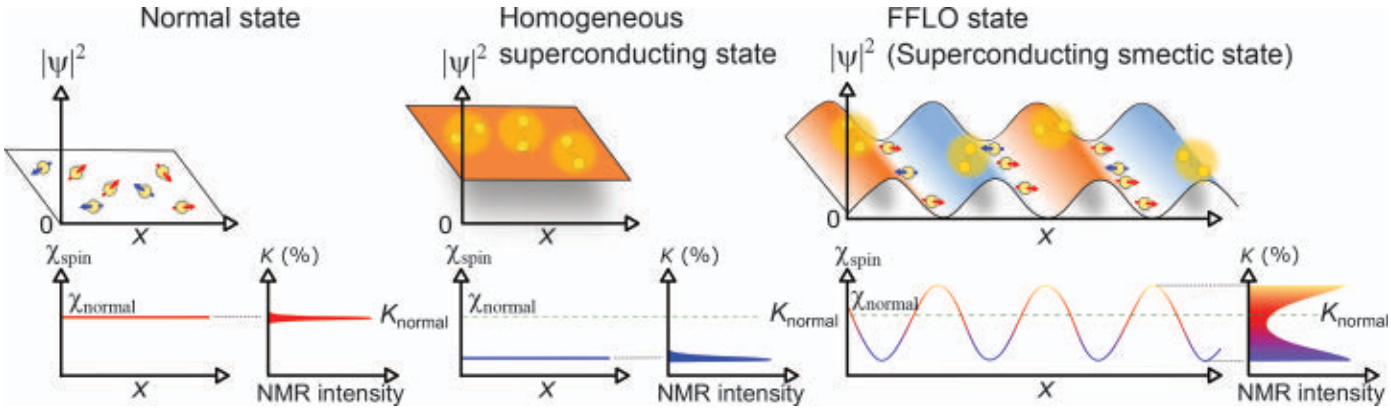


Fig. 1. Schematic comparison among the normal state, homogeneous SC state, and FFLO SC state. (**Left**) In the normal state, conduction electrons having spin 1/2 are polarizable and have a finite spin susceptibility $\chi_{spin} = \chi_{normal}$. This χ_{spin} is detected as a single NMR peak centered at K_{normal}. (**Middle**) In an ordinary spin-singlet superconductor with homogeneous SC order parameter $|\Psi|$, electrons uniformly form spin-less pairs, and thus spin-density is homogeneously reduced from χ_{normal}. (**Right**) By contrast, in the FFLO state the spatial modulation of $|\Psi|$ breaks the translational symmetry, with nodes appearing periodically. Because the adjacent antinodes have opposite order-parameter signs as depicted by the colors on the schematic at top right, quasiparticles induced around the nodes are more spin-polarized than that in the normal state. This results in SC spin smecticity with nontrivial local spin enhancement. The spin smecticity is evidenced by the double-horn NMR spectra, with one peak having a Knight shift larger than K_{normal}.

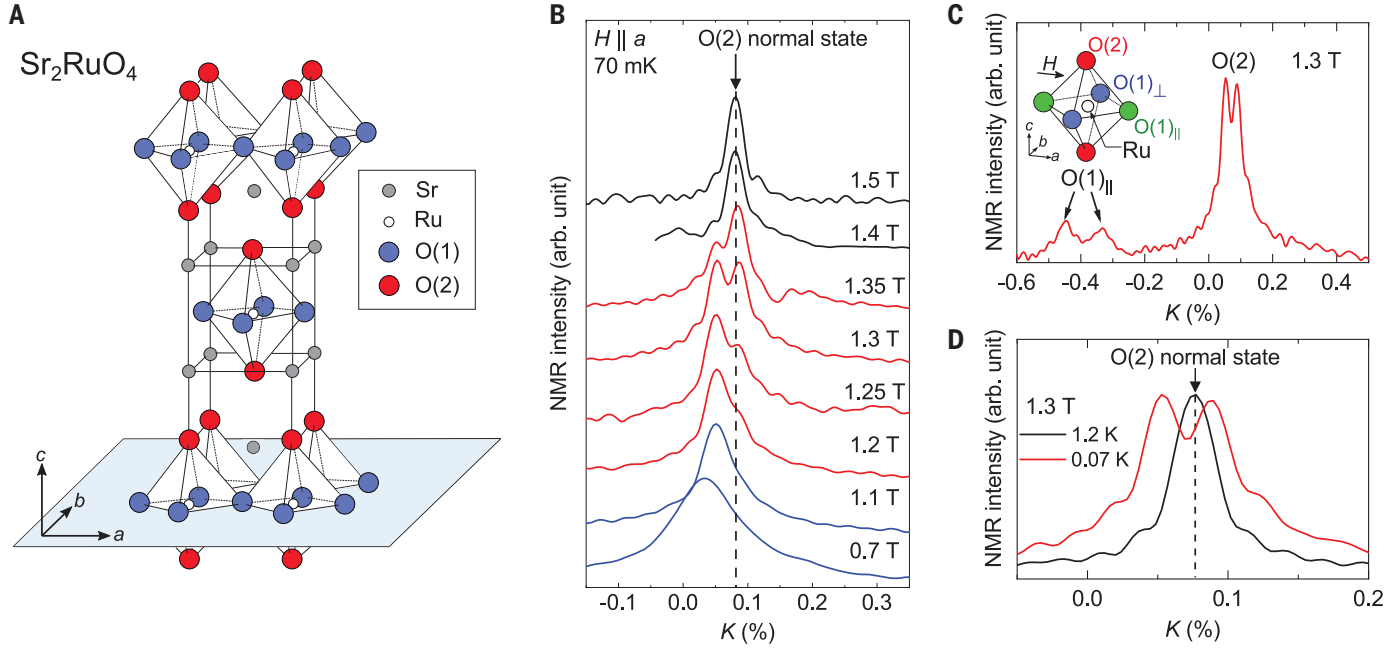


Fig. 2. Double-horn NMR evidence for the FFLO state. (**A**) Crystal structure of Sr_2RuO_4. The blue plane indicates the RuO_2 conduction layer. (**B**) ^{17}O-NMR spectrum at the O(2) site measured at 70 mK and various fields. The spectra in red exhibit the two-peak structure that indicates the FFLO state. The spectra are offset for clarity. (**C**) Comparison of the spectra for the $O(1)_{||}$ and O(2) sites. Both the O(2) site (red circle) and the $O(1)_{||}$ site (green circle), whose Ru-O bonding is parallel to the applied magnetic field, exhibit the characteristic double-horn structure. (**D**) Comparison of O(2)-site spectra at 1.2 K (normal state) and 0.07 K (FFLO state) at 1.3 T. The intensity is normalized with the maximum at each temperature. The right-side peak has a larger Knight shift than that of the normal state, indicating the emergence of a nontrivial spin-dense region.

behave as those for conventional spin-singlet superconductors; the Knight shift at the peak decreases owing to the reduced electron spin susceptibility of the pairs, and the linewidth increases because of the distribution of the local magnetic field in the vortex state. Between these regions, unusual behavior occurs: An additional NMR peak appears starting from 1.2 T and grows with increasing H, resulting in a "double-horn" structure (two peaks having equal intensities) at 1.3 T. A modulation of similar magnitude (~40% of the spin part of the Knight shift) was also observed at the $O(1)_{||}$ site (Fig. 2C). Comparison between O(2) and $O(1)_{||}$ NMR spectra may reveal the orbital dependence of the FFLO state because they hybridize with different sets of Ru-4d orbitals, as discussed in (*27*) (fig. S6). Further, K of the right-side peak is larger than that of the normal-state peak, as shown in Fig. 2D, indicating the realization of a large spin-dense region in the sample. This two-peak spectrum cannot be ascribed to a simple phase separation between the SC and normal states (*27*).

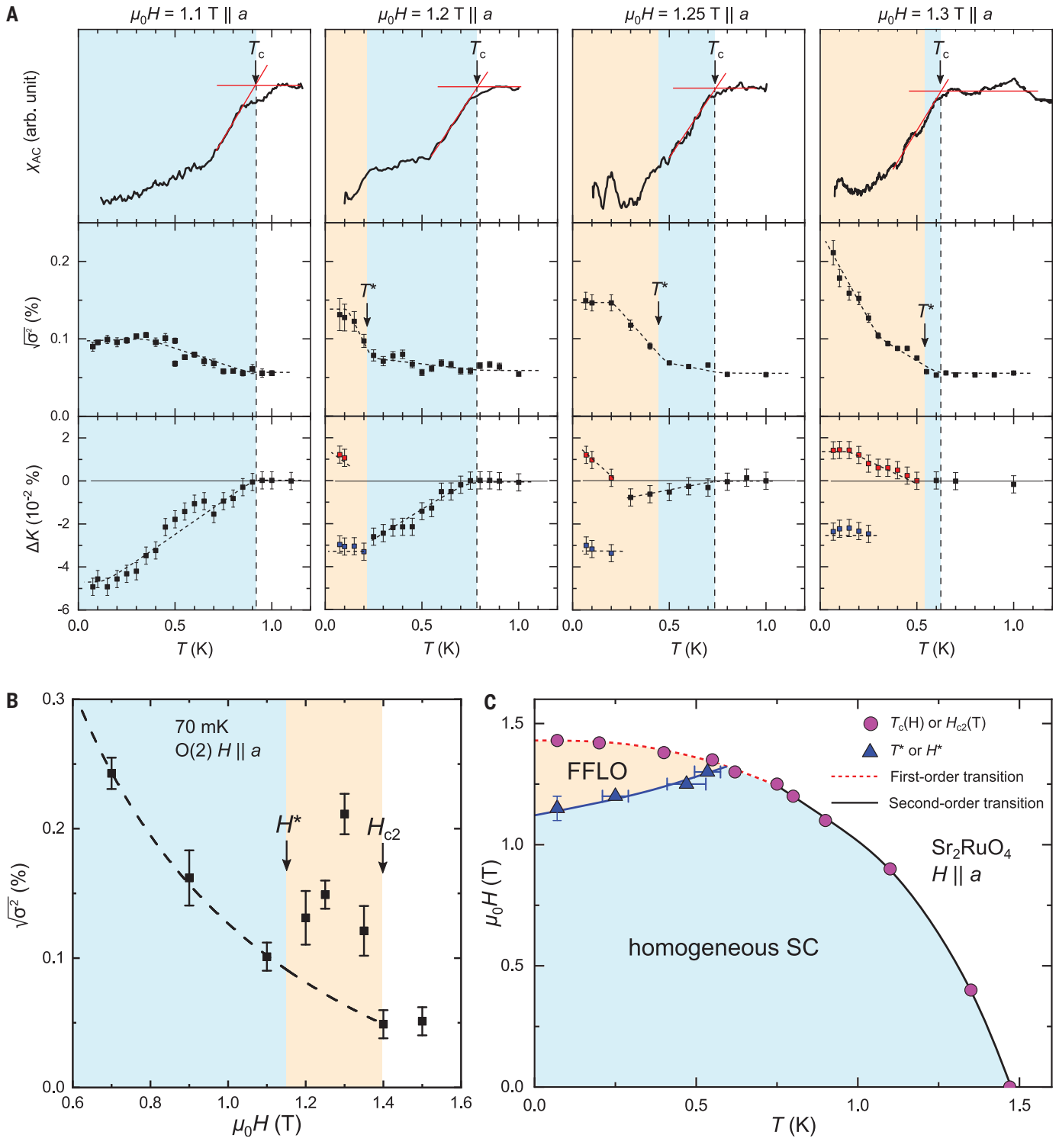


Fig. 3. NMR spectrum change and phase diagram of Sr_2RuO_4. (**A**) (Top) SC shielding signal measured with the frequency around 7.4 MHz. (Middle) Square root of the second moment $\sqrt{\sigma^2}$ of the NMR peaks corresponding to the spectral width at the O(2) site. T^* is defined as the temperature below which $\sqrt{\sigma^2}$ exceeds the ordinary vortex-lattice contribution shown as the dashed curve in (B). (Bottom) Change in the NMR Knight shift from the normal-state value. Dashed lines in the middle and bottom are guides to the eye. The background colors are white for the normal state, orange for the anomalous SC state, and blue for the homogeneous SC state. (**B**) Magnetic-field dependence of $\sqrt{\sigma^2}$ at 70 mK. The dashed line is the extrapolation from low-H behavior assuming vortex-lattice contribution (*27*). (**C**) Phase diagram representing the FFLO region determined from the NMR results. The red dashed curve indicates the FOT determined from the magnetocaloric effect (*26*).

To obtain further information, we investigated the temperature variation of the O(2)-site NMR spectrum as well as the bulk shielding signal to determine T_c. Shown in Fig. 3A are the temperature variation of the ac susceptibility (χ_{AC}); the square root of the second moment ($\sqrt{\sigma^2}$), characterizing the NMR spectral linewidth (*27*) and reflecting the distribution of spin susceptibility; and the Knight-shift change from the normal-state value, measured at various fields. These NMR quantities were determined from the spectra shown in fig. S4. At 1.1 T, the decreasing K (ΔK) and increasing $\sqrt{\sigma^2}$ upon cooling below T_c is the above-mentioned conventional spin-singlet SC behavior. At 1.25 T, an additional peak with a larger Knight shift than the normal-state shift appears only below 0.3 K, which is far below T_c ~ 0.7 K. Moreover, the field dependence of $\sqrt{\sigma^2}$ shown in Fig. 3B exhibits an anomalous enhancement near H_{c2}. A comparison between this behavior and that reported in organic superconductors (*12*) is discussed in (*27*). Such an anomalous $\sqrt{\sigma^2}$ at low temperatures indicates that an inhomogeneous superconductivity emerges in the high-field region.

To determine the boundary of this inhomogeneous state in the *H-T* phase diagram systematically, we focused on $\sqrt{\sigma^2}$, which characterizes the distribution of spin density. Because $\sqrt{\sigma^2}$ exhibits additional rapid increase below a temperature substantially lower than T_c, we define T^* below which $\sqrt{\sigma^2}$ exceeds the value expected from the ordinary behavior stemming from vortex penetration (Fig. 3B, dashed curve) (*27*). T^* characterizes the onset of spectral splitting. As shown in Fig. 3C, T^* increases with increasing H, indicating that the inhomogeneous state becomes more stable when $H_{c2}(0)$ is approached. Thermodynamic measurements on Sr_2RuO_4 have not revealed this phase boundary. However, an anomaly in fourfold in-plane field anisotropy of the specific heat above about 1.2 T, close to H^* in our phase diagram, has been reported (*28*, *29*); this anomaly was interpreted as the occurrence of an extra phase, such as the FFLO phase (*29*).

The double-horn two-peak spectra with one of the peaks having a larger value of K than that of the normal state were observed only in the high-H and low-T region (Fig. 3C). This peak splitting cannot be explained by ordinary phase separation. Instead, such a double-horn structure is naturally explained by the real-space sinusoidal spin modulation with a single wave number, as shown schematically in Fig. 1. Such double-horn NMR spectra have often been observed in spin-density wave states with a sinusoidal-modulated ordered moment (*30*). The crucial difference in the present case is that the spin modulation is formed with superfluid electron pairs. Thus, the present observation provides evidence for SC spin smecticity. For comparison, the Q-phase of $CeCoIn_5$ exhibits a spin modulation much larger than the inhomogeneous phase in Sr_2RuO_4. This difference is consistent with the spin modulation in the $CeCoIn_5$ Q-phase being dominated by magnetic order, whereas that in Sr_2RuO_4 is a result of the SC spin smecticity (*27*).

The observed spin smecticity provides unambiguous evidence for the FFLO state. The distribution of the spin density in an FFLO state has been theoretically calculated (*31*, *32*). The simulated NMR spectra (*31*) for the FFLO state at $H = 0.9\ H_{c2}$ is in good agreement with our double-horn spectrum at 1.3 T (fig. S7). Additionally, the field and temperature dependence of the spectral shape (Fig. 2B and fig. S4) can be understood within the FFLO scenario. The intensity of the peak arising from the nodal region (the right-side peak) in the FFLO state increases abruptly with increasing H and decreasing T owing to the formation of the nodal plane. Moreover, the intensity ratio of the two peaks depends on the FFLO wavelength q and thus on temperature and the magnitude of the applied H. The real-space nodes in the FFLO state can be viewed as topologically protected bound states formed inside the SC gap (*33*, *34*). The π-phase shift of the pair potential across the nodal plane leads to zero-energy bound states with an enhanced density of states. This enhancement makes the spin density at the FFLO nodal plane higher than the value in the normal state.

In the FFLO state, Cooper pairs ($\boldsymbol{k}\uparrow, -\boldsymbol{k}+\boldsymbol{q}\downarrow$) with a finite total momentum $|\boldsymbol{q}| \sim g\mu_B H/\hbar v_F$ are formed between the electrons in the Zeeman-split parts of the Fermi surfaces. Here, g represents the electron-spin g-factor, μ_B is the Bohr magneton, $\hbar$ is Planck's constant h divided by 2π, and v_F is the Fermi velocity. Consequently, the SC order parameter is modulated in the real space as $\Delta(\boldsymbol{r}) = \Delta\cos(\boldsymbol{q}\cdot\boldsymbol{r})$, with the wavelength $\lambda_{FFLO} = 2\pi/|\boldsymbol{q}|$. From the comparison between the NMR spectrum at 1.3 T and the simulation (fig. S7) (*27*, *31*), λ_{FFLO} is estimated to be 30 times longer than the in-plane coherence length ξ, and thus $\lambda_{FFLO} \sim 30\xi \sim 2.0\ \mu m$ by using $\xi \sim 0.066\ \mu m$ of Sr_2RuO_4. This is in good agreement with the crude estimation $\lambda_{FFLO} = 2\pi\hbar v_F/g\mu_B H = 2\pi^2\Delta\xi/g\mu_B H = 29.8\xi = 1.97\ \mu m$, assuming $g = 2$ and $\Delta = 1.76\ k_B T_c$, where k_B is the Boltzmann constant.

As the next step to obtain the definitive evidence of the FFLO state, it is quite important to detect the FFLO phase boundary from other thermodynamic measurements and to image the spatial modulation of the SC order parameter directly, which should be possible with scanning tunneling microscopy. Our study reveals that Sr_2RuO_4 is a most suitable superconductor for studying the FFLO state because the FFLO state is observable under magnetic fields one order of magnitude smaller than those for the other FFLO candidates. This makes the inhomogeneous SC state we observed in Sr_2RuO_4 accessible to various experimental probes.

REFERENCES AND NOTES

1. P. Fulde, R. A. Ferrell, *Phys. Rev.* **135** (3A), A550–A563 (1964).
2. A. I. Larkin, Y. N. Ovchinnikov, *Zh. Eksperim. i Teor. Fiz.* **47**, 1136 (1964).
3. J. J. Kinnunen, J. E. Baarsma, J.-P. Martikainen, P. Törmä, *Rep. Prog. Phys.* **81**, 046401 (2018).
4. A. Sedrakian, J. W. Clark, *Eur. Phys. J. A* **55**, 167 (2019).
5. Y. Matsuda, H. Shimahara, *J. Phys. Soc. Jpn.* **76**, 051005 (2007).
6. R. Lortz *et al*., *Phys. Rev. Lett.* **99**, 187002 (2007).
7. A. Bianchi, R. Movshovich, C. Capan, P. G. Pagliuso, J. L. Sarrao, *Phys. Rev. Lett.* **91**, 187004 (2003).
8. M. Kenzelmann *et al*., *Science* **321**, 1652–1654 (2008).
9. H. Mayaffre *et al*., *Nat. Phys.* **10**, 928–932 (2014).
10. G. Koutroulakis, H. Kühne, J. A. Schlueter, J. Wosnitza, S. E. Brown, *Phys. Rev. Lett.* **116**, 067003 (2016).
11. S. Kitagawa *et al*., *Phys. Rev. Lett.* **121**, 157004 (2018).
12. B. M. Rosemeyer, A. B. Vorontsov, *Phys. Rev. B* **94**, 144501 (2016).
13. Y. Maeno *et al*., *Nature* **372**, 532–534 (1994).
14. A. P. Mackenzie, Y. Maeno, *Rev. Mod. Phys.* **75**, 657–712 (2003).
15. Y. Maeno, S. Kittaka, T. Nomura, S. Yonezawa, K. Ishida, *J. Phys. Soc. Jpn.* **81**, 011009 (2012).
16. A. Pustogow *et al*., *Nature* **574**, 72–75 (2019).
17. K. Ishida, M. Manago, K. Kinjo, Y. Maeno, *J. Phys. Soc. Jpn.* **89**, 034712 (2020).
18. A. N. Petsch *et al*., *Phys. Rev. Lett.* **125**, 217004 (2020).
19. K. Ishida *et al*., *Nature* **396**, 658–660 (1998).
20. A. Steppke *et al*., *Science* **355**, eaaf9398 (2017).
21. V. Grinenko *et al*., *Nat. Phys.* **17**, 748–754 (2021).
22. S. Ghosh *et al*., *Nat. Phys.* **17**, 199–204 (2021).
23. S. Benhabib *et al*., *Nat. Phys.* **17**, 194–198 (2021).
24. S. A. Kivelson *et al*., *npj Quantum Mater.* **43**, 5 (2020).
25. H. G. Suh *et al*., *Phys. Rev. Res.* **2**, 032023 (2020).
26. S. Yonezawa, T. Kajikawa, Y. Maeno, *Phys. Rev. Lett.* **110**, 077003 (2013).
27. Materials and methods are available as supplementary materials.
28. K. Deguchi, Z. Q. Mao, H. Yaguchi, Y. Maeno, *Phys. Rev. Lett.* **92**, 047002 (2004).
29. S. Kittaka *et al*., *J. Phys. Soc. Jpn.* **87**, 093703 (2018).
30. K. Nawa, M. Takigawa, M. Yoshida, K. Yoshimura, *J. Phys. Soc. Jpn.* **82**, 094709 (2013).
31. M. Ichioka, H. Adachi, T. Mizushima, K. Machida, *Phys. Rev. B Condens. Matter Mater. Phys.* **76**, 014503 (2007).
32. K. M. Suzuki, Y. Tsutsumi, N. Nakai, M. Ichioka, K. Machida, *J. Phys. Soc. Jpn.* **80**, 123706 (2011).
33. T. Mizushima, K. Machida, M. Ichioka, *Phys. Rev. Lett.* **95**, 117003 (2005).
34. A. V. Vorontsov, J. A. Sauls, M. J. Graf, *Phys. Rev. B Condens. Matter Mater. Phys.* **72**, 184501 (2005).
35. K. Kinjo, abb0332_figure.zip. *Zenodo* (2022).

ACKNOWLEDGMENTS

The authors thank Y. Yanase, S. Kasahara, Y. Matsuda, S. Kittaka, K. Machida, M. Ichioka, A. P. Mackenzie, and S. E. Brown for valuable discussions. **Funding:** This work was partially supported by the Kyoto University LTM center, Grant-in-Aid for Scientific Research on Innovative Areas from the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) of Japan, and JSPS Core-to-Core program (JPJSCCA20170002). We are also supported by JSPS KAKENHI JP15H05852, JP15K21717, JP17H06136, JP20H00130, JP15K21732, JP15H05745, JP20KK0061, JP19H04696, JP19K14657, and JP20H05158. **Author contributions:** K.I. and Y.M. implemented this project. Z.Q.M., S.Y., and Y.M. synthesized and characterized the samples. K.K., M.M., S.K., and K.I. performed the NMR measurements and analysis. K.K., S.K., K.I., S.Y., and Y.M. drafted the manuscript. All authors contributed to interpret the experimental results and to finalize the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Data supporting this study's findings are available in (*35*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abb0332
Materials and Methods
Supplementary Text
Figs. S1 to S8
Reference (*36*)

8 March 2021; accepted 22 March 2022
10.1126/science.abb0332

QUANTUM SENSING

Atomic-scale quantum sensing based on the ultrafast coherence of an H_2 molecule in an STM cavity

Likun Wang[1], Yunpeng Xia[1], W. Ho[1,2]*

A scanning tunneling microscope (STM) combined with a pump-probe femtosecond terahertz (THz) laser can enable coherence measurements of single molecules. We report THz pump-probe measurements that demonstrate quantum sensing based on a hydrogen (H_2) molecule in the cavity created with an STM tip near a surface. Atomic-scale spatial and femtosecond temporal resolutions were obtained from this quantum coherence. The H_2 acts as a two-level system, with its coherent superposition exhibiting extreme sensitivity to the applied electric field and the underlying atomic composition of the copper nitride (Cu_2N) monolayer islands grown on a Cu(100) surface. We acquired time-resolved images of THz rectification of H_2 over Cu_2N islands for variable pump-probe delay times to visualize the heterogeneity of the chemical environment at sub-angstrom scale.

Quantum sensing and quantum computing, together with other quantum processes, have shown advantages over their classical counterparts (*1*, *2*). Unlike quantum computing, which pursues the long decoherence time in a robust quantum system such as an isolated qubit, quantum sensing capitalizes on the weakness of a quantum system for its high sensitivity to the external environment. Although nitrogen vacancy (NV) centers (*3*–*6*), trapped ions (*7*), and single-electron transistors (*8*, *9*) have been used as quantum sensors, atomic-scale spatial resolution has been impeded by the large size of existing sensors or the limitations of experimental techniques.

The scanning tunneling microscope (STM) offers atomic-scale measurement and control of molecular systems in a surrounding environment that can be characterized by imaging (*10*). A molecular quantum dot has been attached to the STM tip to measure the surface electric potential (*11*). Electron spin resonance sensors based on single magnetic atom–functionalized tips have probed the local magnetic fields produced by atomic and molecular spins (*12*–*16*). The combination of STM and femtosecond lasers has probed the temporal dynamics of molecular motions in the STM cavity with atomic-scale spatial resolution far below the diffraction limit (*17*–*19*).

Here, we studied single H_2 molecules in the cavity defined by the Ag tip and a Cu_2N island grown on the Cu(100) surface using a femtosecond laser, corresponding to terahertz (THz) frequencies, together with a low-temperature STM. By performing THz rectification spectroscopy (TRS) and THz pump-probe measurements, we demonstrate the sensitivity of the coherence of a single H_2 molecule to its immediate environment. The THz pulses in the STM cavity can couple two low-lying states of H_2 in a double-well potential and create a superposition that oscillates periodically with a frequency corresponding to the energy separation of the two states. The damping of this coherent oscillation provides a measure of the decoherence time from the interaction of the two-level system (TLS) with its surrounding environment.

Experiments were performed in a home-built ultrahigh-vacuum (UHV) STM at a base temperature of 9.0 K and with silver (Ag) tips. A femtosecond Ti:sapphire laser with 1-GHz repetition rate was used to generate THz pulses from a plasmonic photoconductive antenna (*20*). The THz pulses were aligned and focused into the cavity through flat silver mirrors and aspheric Tsurupica lenses. The experimental approach is shown schematically in Fig. 1A, and additional details are described in the supplementary materials (figs. S1 and S2).

Hydrogen molecules adsorb on a variety of materials, including metal surfaces (*21*), insulating layers (*22*–*26*), and single molecules (*27*, *28*). To describe a H_2 molecule in the tunnel junction, a TLS in a double-well potential has been widely adopted (*29*, *30*) in which the H_2 molecule switches between two different adsorption configurations (*31*) (Fig. 1B). The population change from the predominant lower state $|a\rangle$ to upper state $|b\rangle$ can be greatly enhanced when the tunneling electrons have sufficient energy to excite H_2 external vibrational or rotational states (*30*) lying above the central barrier of the double-well potential.

With a H_2 molecule weakly trapped inside the STM cavity, the spatial resolution in constant-current topography can be greatly enhanced (*22*). A topographic image of an incommensurate Cu_2N island grown on Cu(100) (*32*) is shown in Fig. 1C, with a close-up image in Fig. 1D. By performing inelastic electron tunneling spectroscopy (IETS) measurements over different sites of the Cu_2N island (Fig. 1E), we confirmed the presence of an H_2 molecule in the cavity (*25*–*27*). The $v = 0 \rightarrow 1$ excitation of the external vibration at ±20 mV and $j = 0 \rightarrow 2$ (para-hydrogen) rotational excitation at ±43 mV did not show resolvable differences for the three high-symmetry positions over Cu_2N.

The spatial variations of the charge distribution on the heteroatomic Cu_2N surface (*33*) were expected to induce a dipole moment in the adsorbed H_2 molecules. Femtosecond THz laser pulses irradiating the STM cavity can excite the TLS, change its population, and facilitate tunneling of H_2 between the two states. To extract the population change from the tunneling current induced by THz irradiation, we implemented atomic-scale rectification spectroscopy, which was initially demonstrated in the microwave frequency range (*34*, *35*). The calibration of the absolute rectification current is shown in fig. S3. A peak voltage of 4.2 mV was derived from THz irradiation by comparing the peak width of IETS and TRS measurements. Further discussion on this weak THz field is given in the supplementary materials (fig. S4). Single-beam TRS shows unchanged vibrational and rotational excitation energies (Fig. 1F) for the same tip-substrate separation and substrate positions as IETS in Fig. 1E. However, the spectral line shapes in Fig. 1, E and F, substantially differ because of population change in the H_2 TLS induced by the THz pulses.

To monitor the temporal evolution of the THz-induced population change, we conducted THz pump-probe measurements over H_2 trapped between the STM tip and the Cu_2N surface. The rectification current was recorded as a function of the time delay τ between two nearly identical THz beams (Fig. 2A). Coherent oscillations with temporal decay and beating were clearly resolved with the tip positioned over different positions of the Cu_2N island. At each of the three high-symmetry positions, a strong peak appeared in the fast Fourier transform (FFT) that corresponded to the main oscillation in the time domain. Additionally, the beating in each delay scan led to satellite peaks in the frequency domain. Similar oscillation was also seen in the spectral intensity of TRS for different delay times, as shown in fig. S5. Additional details and analyses are described in the supplementary materials.

The coherent oscillations of H_2 were sensitive to the position of the tip over the Cu_2N island (Fig. 2A). The description of the Cu_2N lattice is given in fig. S6. No substantial difference in the vibrational and rotational excitation energies could be resolved by IETS and single-beam TRS for the STM tip over the three lateral positions of the Cu_2N layer. In contrast, the THz pump-probe measurements revealed largely distinct frequencies for the three positions (Fig. 2B). For example, an

[1]Department of Physics and Astronomy, University of California, Irvine, CA 92697, USA. [2]Department of Chemistry, University of California, Irvine, CA 92697, USA.
*Corresponding author. Email: wilsonho@uci.edu

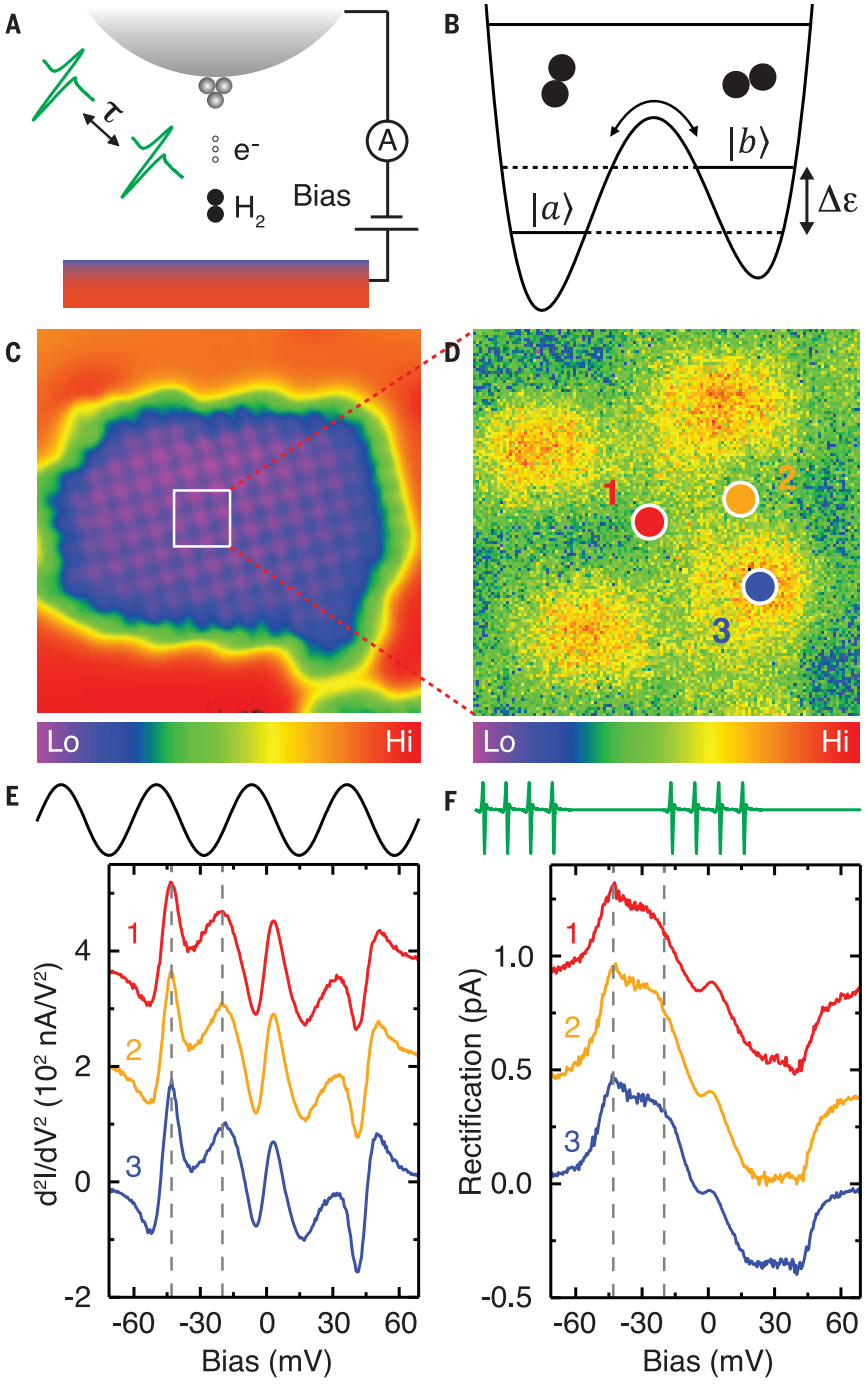


Fig. 1. IETS and TRS of H_2 over different sites of Cu_2N. (**A**) Schematic diagram of THz-STM experimental setup. τ is the temporal delay between the pump-probe THz pulses. (**B**) Asymmetric double-well potential of H_2 molecule in STM cavity. (**C**) Constant-current topography of Cu_2N island acquired with Ag tip; image size 62.0 Å × 62.0 Å. Scanning condition: −20 mV/0.3 nA. (**D**) Zoom-in topographic image of the Cu_2N island, 7.4 Å × 7.4 Å, with red (1), yellow (2), and blue (3) dots over the hollow, bridge, and top sites of the Cu_2N lattice, respectively. (**E**) STM-IETS measurements of H_2 molecule acquired with root-mean-square bias modulation of 3 mV at 263.03 Hz. Tip is positioned at the three different positions labeled in (D) with −20 mV/40 pA setpoint. The black curve in the upper schematic illustrates the sine-wave modulation for IETS. (**F**) TRS measurements of H_2 molecule over the three high-symmetry positions. Single beam of THz pulses is aligned into the junction with the same setpoint and chopped at the same frequency as the STM-IETS for acquiring the spectra. The green curve in the upper schematic illustrates modulation of the single THz beam for TRS. Within half of each square-wave modulation period, a pulse train of ~1.9 million THz pulses is directed into the STM cavity. Dashed lines in (E) and (F) mark the excitation energies of the external vibration and rotation at −20 mV and −43 mV, respectively. All of the spectra in (E) and (F) are offset vertically for clarity.

oscillation frequency shift of 0.11 THz was recorded from position 2 (0.26 THz) to position 3 (0.37 THz), which corresponds to a sensitivity of 58 GHz/Å in the surface plane. The frequency difference was well resolved in the FFT spectrum even though the lateral distance between positions 2 and 3 was only 1.9 Å.

The lattice mismatch between Cu_2N and Cu(100) (*32*) and the asymmetry of the STM tip could induce multiple closely related configurations of H_2 and lift the degeneracy of the TLS energy levels, which led to satellite peaks in the FFT spectrum (Fig. 2B). By fitting multiple sinusoidal functions with an exponential decay term to the coherent oscillations over each position (fig. S7), we could also extract the decoherence time T_2, as shown in the inset of Fig. 2B. Positions 1 and 3 showed similar T_2 around 45 ps, whereas position 2 exhibited a longer time of 80 ps. A set of THz pump-probe measurements with the STM tip over a series of the protrusion sites across the Cu_2N island is shown in fig. S8, illustrating the high spatial sensitivity of the coherent oscillations and surface heterogeneity. In addition, the coherent oscillation over position 1 of fig. S8A exhibited negligible beating and revealed a single peak located at 0.444 THz in the FFT spectrum (fig. S9). The full width at half-maximum (FWHM) of the peak was 4.3 GHz by fitting with a Gaussian peak. Taking the resolving power to be the separation of two peaks by the FWHM, the THz pump-probe measurements could resolve a frequency shift of 4.3 GHz or an energy of 18 μeV.

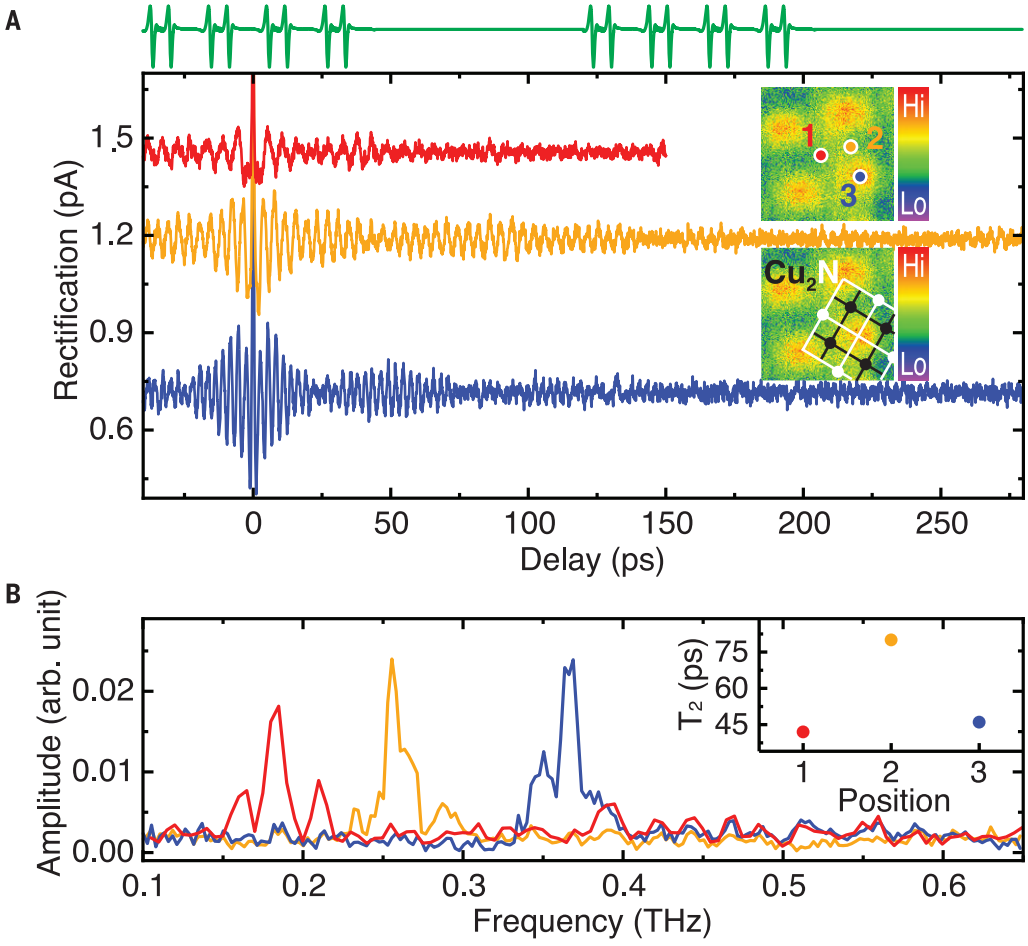


Fig. 2. THz pump-probe measurements of H_2 over different positions of Cu_2N. (**A**) Time-domain measurements of the THz rectification current through H_2 over three lateral positions. The tip is positioned over the correspondingly colored spots with the feedback setpoint −20 mV/40 pA. The feedback is then turned off and sample bias is ramped to −30 mV for the measurements. All of the spectra are offset vertically for clarity. The green curve in the upper schematic illustrates modulation of THz pump-probe beams for TRS. Within half of each square-wave modulation period, a pulse train of ~1.9 million THz pulse pairs is directed into the STM cavity. Inset: Constant-current STM topography of Cu_2N surface, 7.4 Å × 7.4 Å, showing the tip positions (1, 2, and 3) and lines connecting the Cu and N atoms in the two sublattices. (**B**) FFT results of the corresponding measurements in (A). Inset: Decoherence times extracted from fitting of the data shown in (A).

The periodic arrangement of non-neutral Cu and N atoms determines the surface chemical environment distribution of Cu_2N (*36*, *37*). When the tip was placed over the three high-symmetry positions with the same setpoint, the H_2 molecule trapped inside the cavity exhibited different dipole moments caused by the heterogeneous atomic composition of the underlying surface. As a result, the energy separation of the two levels changes, making the H_2 molecule a sensitive probe of the surface chemical environment.

The effect of intermolecular interactions from other adsorbed H_2 molecules was ruled out with the observation of a constant oscillation frequency at various H_2 concentrations (fig. S10). Unlike the IETS and single-beam TRS, which showed no resolvable position-dependent differences on the excitation energies, THz pump-probe measurements probing the TLS coherence showed an advantage in sensing changes in the molecular environment. In addition to the frequency shift, the decoherence time of the oscillations also exhibited spatial dependence at the atomic scale.

The energy difference of the TLS could also be tuned by varying the tip-substrate separation, which was controlled by adjusting the tunneling current at constant sample bias. As the tunneling current increased from 0.02 to 0.2 nA, the tip moved toward the surface by ~1 Å. A series of THz pump-probe measurements at different tip-substrate separations (Fig. 3A) showed that as the tip approached the substrate, the coherent oscillation gradually became weaker. A color map of the measurements (fig. S11B) shows changes in the oscillation period and amplitude as a function of tip-substrate separation.

The frequency from the FFT exponentially increased as the tip-substrate separation decreased (Fig. 3B). A frequency sensitivity of 0.19 THz/Å in the normal direction (tunneling gap) was determined. Both the oscillation amplitude and decoherence time decreased when the tip-substrate separation decreased (Fig. 3A), suggesting a stronger coupling of the H_2 molecule with the junction environment. The variation in the tip-substrate separation across the Cu_2N lattice was only ~0.03 Å, as measured from the line cut in fig. S6F. This height variation contributed ~0.006 THz to the frequency shift observed in Fig. 2B.

To provide further insight into the change in the energy separation of the TLS, we varied the sample bias while monitoring the coherent oscillations (Fig. 3, C and D). The oscillation frequency increased by ~0.2 THz upon changing the sample bias from −50 mV to 40 mV (Fig. 3E). The total electric field experienced by the H_2 molecule trapped in the STM cavity was a combination of the surface electrostatic field and the DC field from the sample bias. As we altered the sample bias without changing the tip position or tip-substrate separation, the

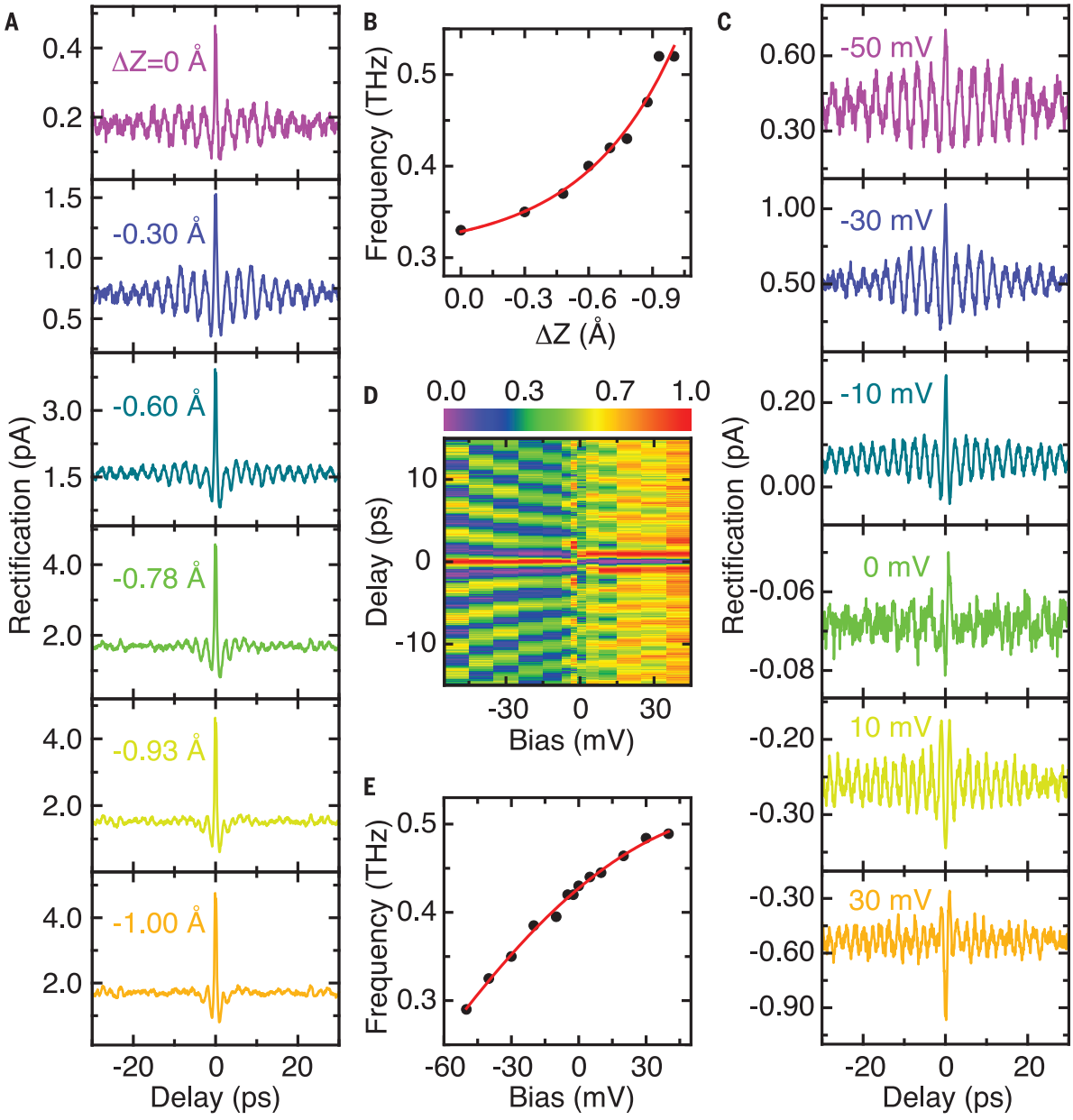


Fig. 3. Coherent oscillation dependency on tip-substrate separation and sample bias. (**A**) THz pump-probe measurements at various tip-substrate separations. The separation is controlled by ramping the tunneling current before turning the feedback off at −20 mV. The bias is then ramped to −30 mV to take the measurements. The colored texts label the corresponding change of tip-substrate separation. (**B**) Oscillation frequency at each change of tip-surface separation from FFT of the time domain measurements in (A); the red line is an exponential fit. (**C**) THz pump-probe measurements at various sample biases. The sample bias is controlled by ramping the bias after turning the feedback off at −20 mV/40 pA. The colored texts label the corresponding sample bias. (**D**) Color map of a series of THz pump-probe measurements, some of which are displayed in (C); the change of the coherent oscillation as a function of sample bias can be visualized. For clarity, the intensity of all of the spectra is normalized positively to the color palette with range (0, 1). (**E**) Oscillation frequency at each bias from FFT of the time domain measurements in (C); the red line is a fitting of the data with function described in the supplementary materials.

DC field varied while the surface electrostatic field remained the same. Because the energy levels shifted under external field through the Stark effect, the energy separation of the TLS changed. Analysis and fitting of the data in Fig. 3E yielded a surface electrostatic field of 43 mV/Å pointing toward the surface. In addition, the dipole moment difference of the two states along this field was extracted and found to be ~0.6 debye. Details of the fitting are described in the supplementary materials. Besides varying the sample bias, the coherent oscillation frequency could also be shifted for different tips under the same sample bias (fig. S12). Effectively, the DC fields varied with tip structure from different tips, which modified the total electric field acting on the H_2 and thus changed its oscillation frequency.

We acquired THz rectification images in real space with atomic-scale spatial and femtosecond-scale temporal resolutions. The spectroscopic features of the TLS exhibited temporal evolution that sensitively depended on the molecule's surrounding environment in the surface plane. As shown in Fig. 4, a series of constant-height THz rectification imaging was recorded at different time delays between pump-probe

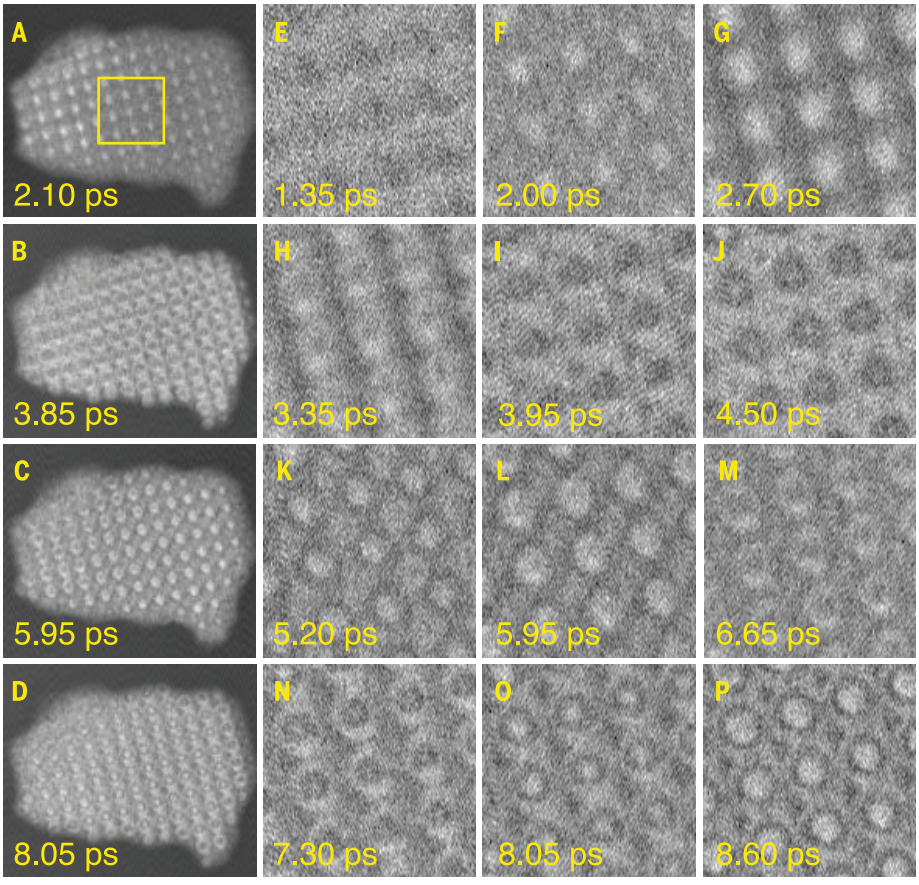


Fig. 4. Temporal snapshots of THz rectification imaging of Cu_2N surface at various delay times. (**A** to **D**) Rectification imaging of the entire Cu_2N island at four different delays. For each pixel in the images, tip is positioned at setpoint −20 mV/40 pA, and then feedback is turned off and bias is ramped to −30 mV. Feedback is turned on before and during the tip movement to the next pixel. All images are 53.2 Å × 44.9 Å and 266 × 224 pixels. (**E** to **P**) Series of zoom-in rectification images at various delays. All images are 13.8 Å × 13.8 Å and 137 × 137 pixels.

pulses. These images provided temporal snapshots of the spatial variations of the H_2 TLS coherence caused by the heterogeneity of the surface. Four successive images of the whole Cu_2N island (Fig. 4, A to D) revealed the wavelike evolution of the "ripples" from the upper left to the bottom right of the island.

A series of images (Fig. 4, E to P) of the yellow square window in Fig. 4A revealed the alternation of the ripple features between bright and dark, or high and low rectification current, as a function of the time delay. Two sequences of temporal snapshots were composed (movies S1 and S2). The properties of the coherent oscillations were highly sensitive to the lateral position of the STM tip (Fig. 2 and fig. S8). In rectification imaging at a chosen delay, the tip was located over different positions of the surface. The oscillation frequency shifted spatially because the surface chemical environment experienced by the H_2 TLS varies over different positions of the surface. The decoherence time also changed as a response to the local environment sensed by the H_2. As a result, the spatially resolved rectification imaging at different delay times effectively revealed the surface chemical environment distribution of the Cu_2N island.

The extreme sensitivity of the H_2 TLS coherent oscillation to the applied electric field and the underlying surface chemical environment heralds the application of the H_2 molecule in the STM cavity for extreme quantum sensing. Relative to other quantum sensors such as NV centers in diamond, the H_2 coherent sensor in the STM cavity provides simultaneous atomic-scale spatial and femtosecond temporal resolutions with GHz energy discrimination. H_2 molecules have been found to be trapped over a variety of surfaces, atoms, and molecules (*22*, *24*, *27*). THz pump-probe measurements with a H_2 molecule in the STM cavity can be used to probe the electrostatic field and potential energy surface of the sample. The ability to measure and control coherent oscillations of the H_2 TLS in different environments opens a route for quantum sensing of a broad range of systems, from single atoms to heteroatomic molecules and solid surfaces.

REFERENCES AND NOTES

1. C. L. Degen, F. Reinhard, P. Cappellaro, *Rev. Mod. Phys.* **89**, 035002 (2017).
2. F. Troiani, A. Ghirri, M. G. A. Paris, C. Bonizzoni, M. Affronte, *J. Magn. Magn. Mater.* **491**, 165534 (2019).
3. M. S. Grinolds *et al.*, *Nat. Phys.* **7**, 687–692 (2011).
4. N. Aslam *et al.*, *Science* **357**, 67–71 (2017).
5. P. Maletinsky *et al.*, *Nat. Nanotechnol.* **7**, 320–324 (2012).
6. K. Bian *et al.*, *Nat. Commun.* **12**, 2457 (2021).
7. M. J. Biercuk, H. Uys, J. W. Britton, A. P. VanDevender, J. J. Bollinger, *Nat. Nanotechnol.* **5**, 646–650 (2010).
8. M. J. Yoo *et al.*, *Science* **276**, 579–582 (1997).
9. J. Martin *et al.*, *Nat. Phys.* **4**, 144–148 (2008).
10. G. Czap *et al.*, *Science* **364**, 670–673 (2019).
11. C. Wagner *et al.*, *Nat. Mater.* **18**, 853–859 (2019).
12. X. Zhang *et al.*, *Nat. Chem.* **14**, 59–65 (2022).
13. P. Willke *et al.*, *Sci. Adv.* **4**, eaaq1543 (2018).
14. T. Choi *et al.*, *Nat. Nanotechnol.* **12**, 420–424 (2017).
15. K. Yang *et al.*, *Science* **366**, 509–512 (2019).
16. S. Baumann *et al.*, *Science* **350**, 417–420 (2015).
17. T. L. Cocker, D. Peller, P. Yu, J. Repp, R. Huber, *Nature* **539**, 263–267 (2016).
18. S. Li, S. Chen, J. Li, R. Wu, W. Ho, *Phys. Rev. Lett.* **119**, 176002 (2017).
19. D. Peller *et al.*, *Nature* **585**, 58–62 (2020).
20. N. T. Yardimci, S. H. Yang, C. W. Berry, M. Jarrahi, *IEEE Trans. Terahertz Sci. Technol.* **5**, 223–229 (2015).
21. J. A. Gupta, C. P. Lutz, A. J. Heinrich, D. M. Eigler, *Phys. Rev. B* **71**, 115416 (2005).
22. S. Liu, A. Shiotari, D. Baugh, M. Wolf, T. Kumagai, *Phys. Rev. B* **97**, 195417 (2018).
23. P. Chen, Z. Xiong, J. Luo, J. Lin, K. L. Tan, *Nature* **420**, 302–304 (2002).
24. F. D. Natterer, F. Patthey, H. Brune, *ACS Nano* **8**, 7099–7105 (2014).
25. F. D. Natterer, F. Patthey, H. Brune, *Phys. Rev. Lett.* **111**, 175303 (2013).
26. S. Li *et al.*, *Phys. Rev. Lett.* **111**, 146102 (2013).
27. S. Li *et al.*, *Phys. Rev. Lett.* **114**, 206101 (2015).
28. J. Jung, S. Nam, C. Wolf, A. J. Heinrich, J. Chae, *RSC Adv.* **11**, 6240–6245 (2021).
29. A. Halbritter, P. Makk, S. Csonka, G. Mihály, *Phys. Rev. B* **77**, 075402 (2008).
30. W. H. A. Thijssen, D. Djukic, A. F. Otte, R. H. Bremmer, J. M. van Ruitenbeek, *Phys. Rev. Lett.* **97**, 226806 (2006).
31. H. Wang *et al.*, *J. Phys. Chem. Lett.* **6**, 3453–3457 (2015).
32. T. Choi, C. D. Ruggiero, J. A. Gupta, *Phys. Rev. B* **78**, 035430 (2008).
33. C. F. Hirjibehedin *et al.*, *Science* **317**, 1199–1203 (2007).
34. J. Lee, X. Tu, W. Ho, *Nano Lett.* **5**, 2613–2617 (2005).
35. X. W. Tu, J. H. Lee, W. Ho, *J. Chem. Phys.* **124**, 021105 (2006).
36. M. Schneiderbauer, M. Emmrich, A. J. Weymouth, F. J. Giessibl, *Phys. Rev. Lett.* **112**, 166102 (2014).
37. K. T. Crampton, J. Lee, V. A. Apkarian, *ACS Nano* **13**, 6363–6371 (2019).

ACKNOWLEDGMENTS

The plasmonic photoconductive antenna was fabricated by M. Jarrahi's group at the University of California, Los Angeles. **Funding:** Supported by US Department of Energy, Office of Basic Energy Sciences, award DE-SC0019448. **Author contributions:** Experiment, analyses, and manuscript preparation were jointly carried out by L.W., Y.X., and W.H. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All the data are available in the main text and the supplementary materials.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abn9220
Materials and Methods
Figs. S1 to S12
Movies S1 and S2

30 December 2021; resubmitted 11 February 2022
Accepted 25 March 2022
10.1126/science.abn9220

OPTOELECTRONICS

Optical absorption of interlayer excitons in transition-metal dichalcogenide heterostructures

Elyse Barré[1,2], Ouri Karni[1,3], Erfu Liu[4], Aidan L. O'Beirne[1,5], Xueqi Chen[5], Henrique B. Ribeiro[1], Leo Yu[3], Bumho Kim[6], Kenji Watanabe[7], Takashi Taniguchi[8], Katayun Barmak[9], Chun Hung Lui[4], Sivan Refaely-Abramson[10], Felipe H. da Jornada[11], Tony F. Heinz[1,3]*

Interlayer excitons, electron-hole pairs bound across two monolayer van der Waals semiconductors, offer promising electrical tunability and localizability. Because such excitons display weak electron-hole overlap, most studies have examined only the lowest-energy excitons through photoluminescence. We directly measured the dielectric response of interlayer excitons, which we accessed using their static electric dipole moment. We thereby determined an intrinsic radiative lifetime of 0.40 nanoseconds for the lowest direct-gap interlayer exciton in a tungsten diselenide/molybdenum diselenide heterostructure. We found that differences in electric field and twist angle induced trends in exciton transition strengths and energies, which could be related to wave function overlap, moiré confinement, and atomic reconstruction. Through comparison with photoluminescence spectra, this study identifies a momentum-indirect emission mechanism. Characterization of the absorption is key for applications relying on light-matter interactions.

The dielectric function is one of the key material characteristics that links fundamental structure and device functionality. It depends nontrivially on the electronic band structure and many-body interactions in a material and is essential for the design of photonic and optoelectronic applications (*1*). In two-dimensional semiconducting monolayers (1L) of transition-metal dichalcogenides (TMDCs), the dielectric function is dominated by resonances associated with strongly bound excitons—correlated electron-hole pairs—arising from the enhanced Coulomb interactions in these materials (*2*). The contribution of excitons to the dielectric function has typically been characterized by their absorption spectra through reflection contrast, $\Delta R/R$ (*3*). In parallel, many excitonic species of various spin and momentum configurations and multi-excitonic states have been identified using photoluminescence (PL) measurements (*3*), which, however, do not permit a determination of the material's dielectric function.

[1]SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA. [2]Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA. [3]Department of Applied Physics, Stanford University, Stanford, CA 94305, USA. [4]Department of Physics and Astronomy, University of California, Riverside, CA 92521, USA. [5]Department of Physics, Stanford University, Stanford, CA 94305, USA. [6]Department of Mechanical Engineering, Columbia University, New York, NY 10027, USA. [7]Research Center for Functional Materials, National Institute for Materials Science, 1-1 Namiki, Tsukuba, Ibaraki 305-0044, Japan. [8]International Center for Materials Nanoarchitectonics, National Institute for Materials Science, 1-1 Namiki, Tsukuba, Ibaraki 305-0044, Japan. [9]Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027, USA. [10]Department of Molecular Chemistry and Materials Science, Weizmann Institute of Science, Rehovot 7610001, Israel. [11]Department of Materials Science and Engineering, Stanford University, Stanford, CA 94305, USA.
*Corresponding author. Email: tony.heinz@stanford.edu

In TMDC heterobilayers—stacks of two different 1L TMDCs—PL has been used to investigate interlayer excitons (ILXs), whose electron and hole constituents reside in opposite layers (*4*). ILXs have demonstrated strong electrical tunability (*5*, *6*) and a rich variety of confined states originating from the periodic potential imposed by the moiré superlattice (*7*, *8*)—that is, the spatially varying atomic configuration between the two layers imposed by lattice and twist-angle mismatch. Because their electrons and holes have little wave function overlap, the ILX absorption is easily masked by the large intralayer absorption, so a direct determination of the ILX dielectric response has remained elusive. As a result, many ambiguities about the nature of ILXs have persisted that are important both for fundamental understanding and for their future use in optical systems: their absorption strength, their momentum-space configuration, their intrinsic radiative lifetime, and the influence of moiré modulation and reconstruction.

We report a direct measurement of the optical absorption of ILX states in the prototypical TMDC heterobilayer of $WSe_2/MoSe_2$ using electromodulation spectroscopy (*1*). This enables us to characterize the ILX contribution to the dielectric function of the material. We focus on H-stacked (60°) heterobilayers, which were previously reported to host a diverse set of ILX states (*9*, *10*). The $WSe_2/MoSe_2$ heterobilayers are encapsulated in hexagonal boron nitride (hBN) and equipped with back- and top-graphite gates (Fig. 1A). By applying an appropriately balanced sinusoidally varying bias voltage to the gates, we induce an alternating electric field F on the heterobilayer with negligible charging (*11*). Because of the finite static electric-dipole moment p of an ILX, its energy E_{ILX} experiences a modulation proportional to the applied field F in the sample: $E_{ILX} = E_0 + pF$ (*5*, *6*). This leads in turn to a modulation of the dielectric function of the heterobilayer, which we record through the resulting modulation of the reflectivity of monochromatic light measured using lock-in detection. To obtain a full spectrum, we tune the probe wavelength across the desired spectral range. By comparing the reflectance with and without modulation and dividing by the electric field amplitude, the fractional change in reflectance with electric field $R_F = (1/R)\partial R/\partial F$ is determined experimentally. For a given excitonic resonance, we then have $R_F = (1/R)(\partial R/\partial E_{ILX})p$, where we determine p by collecting spectra for different dc values of F. We relate R and its derivative $\partial R/\partial E_{ILX}$ to the dielectric function, ε, using a solution to Maxwell's equations, implemented with transfer matrices, for the stacked experimental structure (*11*). We note that R is not modified meaningfully by the ILX resonances because they have small oscillator strength and thus $\partial R/\partial E_{ILX}$ imparts its shape to R_F. To confirm the origin of R_F as a field-induced shift in ILX resonance, we tested the response to intentional charge modulation and found no measurable signal (fig. S2).

Figure 1D presents the ILX absorption spectrum obtained from R_F for a sample with twist angle of 60° ± 0.2° (*11*) compared with the intralayer absorption spectrum from a conventional white-light reflection contrast measurement, $\Delta R/R$. The corresponding underlying measurements of R_F and $\Delta R/R$ are shown in Fig. 1, B and C. These results yield ILX oscillator strengths that are three to four orders of magnitude smaller than those of the intralayer resonances (tables S2 and S7).

In the lower-energy region of R_F, two closely spaced features can be seen, separated by 18 meV. These peaks have previously been observed in helicity-resolved and magneto-PL spectra (*10*, *12*) and have been assigned to 1s spin-antialigned ($a_{\uparrow\downarrow}$) and spin-aligned ($a_{\uparrow\uparrow}$) ILX transitions between the band edges at the K points ($K \rightarrow K$), although PL measurements can also exhibit momentum-indirect transitions (*13*). Our data confirm the momentum-direct assignment, because momentum-indirect excitons are expected to have a much weaker oscillator strength than direct transitions, and thus will be weakly visible in absorption spectra. Unlike in PL measurements, here we can quantify the oscillator strengths of the observed resonances. Surprisingly, the oscillator strength of the spin-antialigned peak, $a_{\uparrow\downarrow}$, is one-fourth as strong as the spin-aligned $a_{\uparrow\uparrow}$ feature. A similar ratio is reproduced in a second H-stacked sample (table S3). This finding is in agreement with our theoretical calculations (table S8) and contrasts sharply with the large oscillator strength difference between spin-split excitons in 1L TMDCs (*14*). This

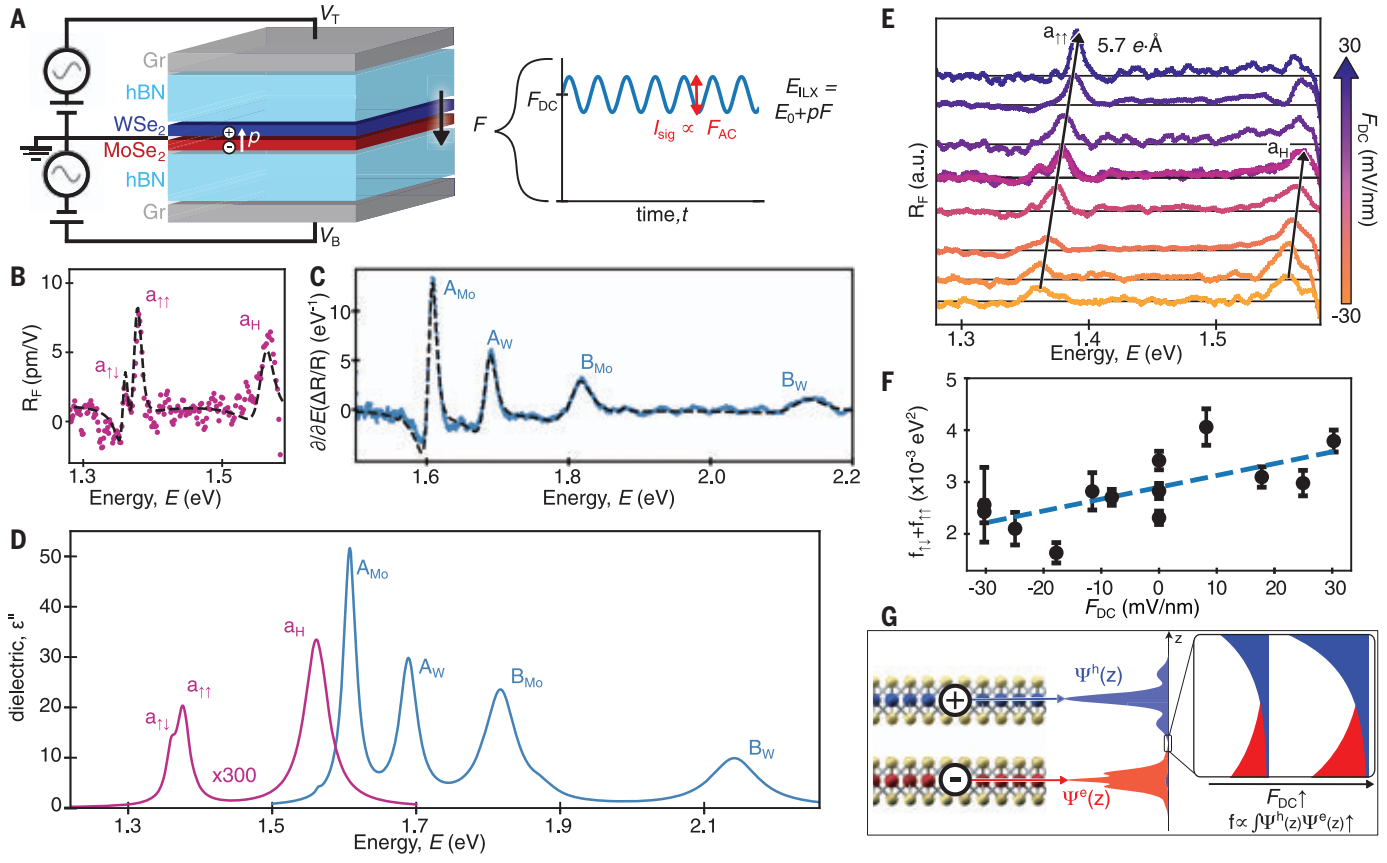


Fig. 1. Absorption spectrum and electric-field dependence of a 60° ± 0.2° aligned sample at 20 K. (**A**) Diagram of an encapsulated, dual-gated $WSe_2/MoSe_2$ heterostructure and the strategy of using an applied electric field, F, to measure the absorption of interlayer excitons through the fractional change in reflectivity with electric field R_F, which is a normalized version of the modulating signal I_{sig}. (**B**) R_F (with F_{AC} = 3.1 ± 0.3 mV_p/nm), shown by magenta dots, and its fit (dashed black line). (**C**) The first derivative of the reflection contrast $\Delta R/R$ versus energy (blue continuous line) and its fit (black dashed line). $A_{Mo(W)}$ and $B_{Mo(W)}$ are the spin-split excitons associated with intralayer transitions at the K point of $MoSe_2$ (WSe_2). (**D**) The imaginary dielectric function extracted from R_F (magenta) and from $\Delta R/R$ (blue). (**E**) R_F as a function of F_{DC}. The inferred static dipole moments are $p_{\uparrow\downarrow}$ = 6.2 ± 0.6 e·Å, $p_{\uparrow\uparrow}$ = 5.7 ± 0.5 e·Å, and p_H = 2.6 ± 0.6 e·Å. (**F**) The increase in summed oscillator strength of $a_{\uparrow\uparrow}$ and $a_{\uparrow\downarrow}$ as a function of F_{DC} with a fitted linear trend (dashed blue line). Error bars denote fitting variance. (**G**) Left: The out-of-plane component of the computed DFT wave functions for the conduction [$\Psi^e(z)$] and valence [$\Psi^h(z)$] states at the K point in reciprocal space. Right: An illustration of the increase in wave function overlap with F_{DC} (right).

distinction can be understood as a consequence of the mirror symmetry breaking along the out-of-plane direction in the heterobilayers (*15*).

The oscillator strengths measured by modulation spectroscopy determine the intrinsic radiative recombination rates of these ILX states (*11*). The corresponding vacuum radiative lifetimes, τ_0 = 400 ± 60 and 105 ± 10 ps, are shorter than the previously reported emission lifetimes, τ_{em}, measured through time-resolved PL, which range from ~1 to 500 ns (*5*, *13*, *16*). The increased emission lifetime reflects the fact that the majority of photoexcited excitons lie outside the light cone as they relax from the initial excited state to the radiative state (*11*). The measured emission time is consequently sensitive to the details of the excitation conditions and relaxation pathways (including nonradiative channels).

Further analyzing the series of resonances revealed by modulation spectroscopy, we used the dc electric field dependence to explore the electron and hole layer localization. Figure 1E displays the evolution of R_F with dc electric field F_{DC}. Peaks $a_{\uparrow\downarrow}$ and $a_{\uparrow\uparrow}$ are found to have dipole moments $p_{\uparrow\downarrow}$ = 6.2 ± 0.6 e·Å and $p_{\uparrow\uparrow}$ = 5.7 ± 0.5 e·Å, respectively. These values are consistent with the picture of electron and hole states whose wave functions are localized in opposite layers and thus are separated by the Mo-W interatomic distance of ~6 Å, in agreement with the density functional theory (DFT) predictions described below. In addition to $a_{\uparrow\downarrow}$ and $a_{\uparrow\uparrow}$, we measured a third, higher-energy peak denoted by a_H (Fig. 1D) with a smaller electric dipole moment of p_H = 2.6 ± 0.6 e·Å. There are many possible candidates for high-energy ILX states, including Rydberg-like states (see table S8 for calculations). However, such high-energy ILX states are expected to have small oscillator strengths relative to the low-energy ILXs. Thus, the increased oscillator strength and the reduced dipole moment can be attributed to a mixing of such high-energy states with the $MoSe_2$ A exciton A_{Mo}, in the fashion of hybridized inter/intralayer excitons recently reported in some bilayer systems (*17*, *18*). As a result of this inter/intralayer exciton mixing, R_F detects several features in the intralayer exciton range (*11*).

The absorption peaks, $a_{\uparrow\uparrow}$ and $a_{\uparrow\downarrow}$, in Fig. 1E become more pronounced under positive F_{DC}. Their oscillator strengths almost double over a range ΔF_{DC} = 60 mV/nm (Fig. 1F). We attribute this enhancement to an increase in electron-hole wave function overlap with electric field, as illustrated in Fig. 1G. This mechanism of enhancing oscillator strength was originally proposed in (*5*) to support an experimentally observed decrease in τ_{em} with increasing F_{DC}. Our measurements of $f(F_{DC})$ closely match this previously reported trend of $1/[\tau_{em}(F_{DC})] \propto f(F_{DC})$.

An important concern in these systems is the influence of the moiré pattern on the properties of the ILX. We examined this issue by investigating the trends in the ILX absorption features with twist angle for alignment near 60°. For the different samples

studied, the energies of the spin-aligned $a_{\uparrow\uparrow}$ resonance generally increase with twist angle (Fig. 2A). To explain the energy shift, we first consider the influence of the twist-dependent momentum mismatch between the conduction and valence band edges, ΔK_θ, as shown schematically in Fig. 2C. The energy of the ILX band intersecting the light cone, $E(\Delta K_\theta)$, can be calculated using the measured electron, hole, and ILX effective masses: $m_e = 0.8m_0$ in $MoSe_2$ (*19*), $m_h = 0.4m_0$ in WSe_2 (*20*), and $M = 1.2m_0$ (*21*). However, this effect alone cannot explain the magnitude of the observed energy shift. We therefore considered the contribution of the moiré potential that spatially confines the ILX and changes its band dispersion. We used a continuum Hamiltonian approach in which the potential depth is a fitting parameter (*11*). For a moiré potential well depth of 120 meV, we predict a variation in ILX transition energy with twist angle that is in reasonable agreement with our experimental data. The inferred well depth is twice that predicted by first-principles GW plus Bethe-Salpeter equation (BSE) calculations (*22*). Similar discrepancies between the moiré potential from first-principles calculations and that inferred from experiment have been reported frequently in the past (*7*, *23*, *24*). The potential $V(r)$ for $\Delta\theta = 56°$ and 58° is shown in Fig. 2, D and E, respectively, and overlaid with the corresponding $E_{moiré}$ in Fig. 2, F and G. Although absorption measurements are more robust against the influence of defects than PL data, which can be dominated by the lowest-energy states, they will still be influenced by possible strain. Strain can modify both the band energies and the size of the moiré unit cell (*24*)—factors that contribute to the variation in experimental data seen in Fig. 2A.

With respect to the variation of transition strength with twist angle, we found, as shown in Fig. 2B, that the ILX oscillator strength (summed over the $a_{\uparrow\downarrow}$ and $a_{\uparrow\uparrow}$ transitions) decreases with increasing crystallographic misalignment (i.e., for twist angles farther from 60°). Because the absorption measurements examine direct transitions, we cannot attribute this effect simply to an increase in momentum mismatch. We must rather invoke the effect of the moiré pattern. This occurs through two separate mechanisms: (i) The layer spacing increases with increasing misalignment (*25*), reducing the electron-hole wave function overlap and the ILX oscillator strength. (ii) The lattice reconstruction is less dominant for smaller moiré periods (*26*). In well-aligned H-stacked $WSe_2/MoSe_2$, this reconstruction favors the H_h^h atomic configuration (fig. S4), which has the strongest optical transitions (*22*). For less aligned samples, the H_h^h region covers a smaller fraction of the interface, reducing the overall transition strength (*11*). The second mechanism alone can explain much of the observed trend in ILX oscillator strength, as shown in Fig. 2B.

Finally, we addressed the long-standing issue of whether ILX emission for the $WSe_2/MoSe_2$ system is dominated by momentum-direct ($K \rightarrow K$) transitions (*4*, *8*, *10*) or by momentum-indirect ($K \rightarrow \Lambda$) transitions (*9*, *13*, *27*). On the theoretical side, we calculated the band structure and ILX absorption spectrum for the H_h^h lattice arrangement that dominates the moiré unit cell for our well-aligned sample. Our computed GW band structure predicts that the conduction band minimum of the aligned heterobilayer lies at the Λ point (Fig. 3A). Correspondingly, the lowest excitation energy calculated using the GW-BSE formalism (Fig. 3B) is primarily composed of the momentum-indirect electron-hole transition $K \rightarrow \Lambda$. Because the $K \rightarrow \Lambda$ transition is a second-order process involving a phonon, we expect it to be present in the emission spectrum but not observable in absorption. The predicted static dipole moments associated with the momentum-indirect and -direct excitons are shown in Fig. 3C; they are compared to the experimentally measured ILX static dipole moments for a well-aligned heterobilayer in Fig. 3D, whose zero-field PL and absorption are shown in Fig. 3E. We found good agreement between calculation and experiment for the energies and the ratio of dipole moments between the absorption and PL peaks, thus linking the absorption to $K \rightarrow K$ transitions and the PL to the $K \rightarrow \Lambda$ transition. Further supporting this identification were measurements of the dependence of the ILX PL on excitation power and temperature (Fig. 3, F and G). These studies revealed the emergence of a higher-energy emission peak, L2, that matches the energy of the ILX absorption feature, $a_{\uparrow\downarrow}$, at elevated temperatures and high excitation powers. A thermal-activation model (Fig. 3G, inset) allows us to infer an energy difference between indirect and direct

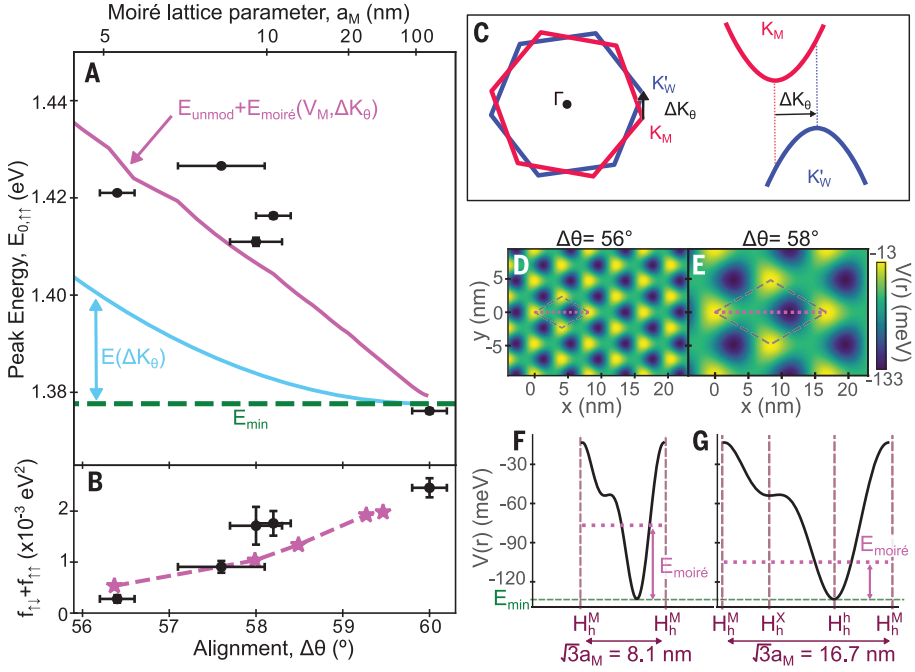


Fig. 2. Twist angle dependence and moiré pattern effects. (**A**) The experimentally measured interlayer spin-aligned ($a_{\uparrow\uparrow}$) exciton energies (black dots; error bars denote fitting variance) compared with theoretically predicted energy shifts. The minimum ILX energy within the moiré, E_{min} (dashed green line), is estimated to be consistent with $E_{0,\uparrow\uparrow}$ at 60°. The cyan curve shows the energy offset for the lowest optically allowed ILX considering only the twist angle–dependent momentum mismatch between the *K* valleys of the constituent materials, ΔK_θ, illustrated in (C). The magenta line shows the energy of the lowest optically allowed ILX considering moiré confinement effects from a moiré potential well depth of 120 meV (defined by the energy at H_h^h to H_h^M), as illustrated in (D) to (G). (**B**) Experimentally measured oscillator strength (black dots; error bars denote fitting variance) versus twist angle. The dashed and starred magenta line shows the expected change in oscillator strength from the relative area of H_h^h to the full moiré unit cell area under reconstruction. (**C**) Diagram showing the origins of ΔK_θ, from momentum offset between the band edges. (**D** and **E**) The real-space moiré potential modulation $V(r)$ for alignments of $\Delta\theta = 56°$ and 58° for the potential defined by the parameters $V_M = 11.76$ meV and $\phi = 139.1°$ (*11*). The moiré unit cell is marked with a brown dashed line; a magenta dashed line through the middle marks the trace for (F) and (G). (**F** and **G**) Moiré potential along a line cut through the moiré unit cell, highlighting the energy of the lowest confined and optically active excitonic state (magenta dashed lines).

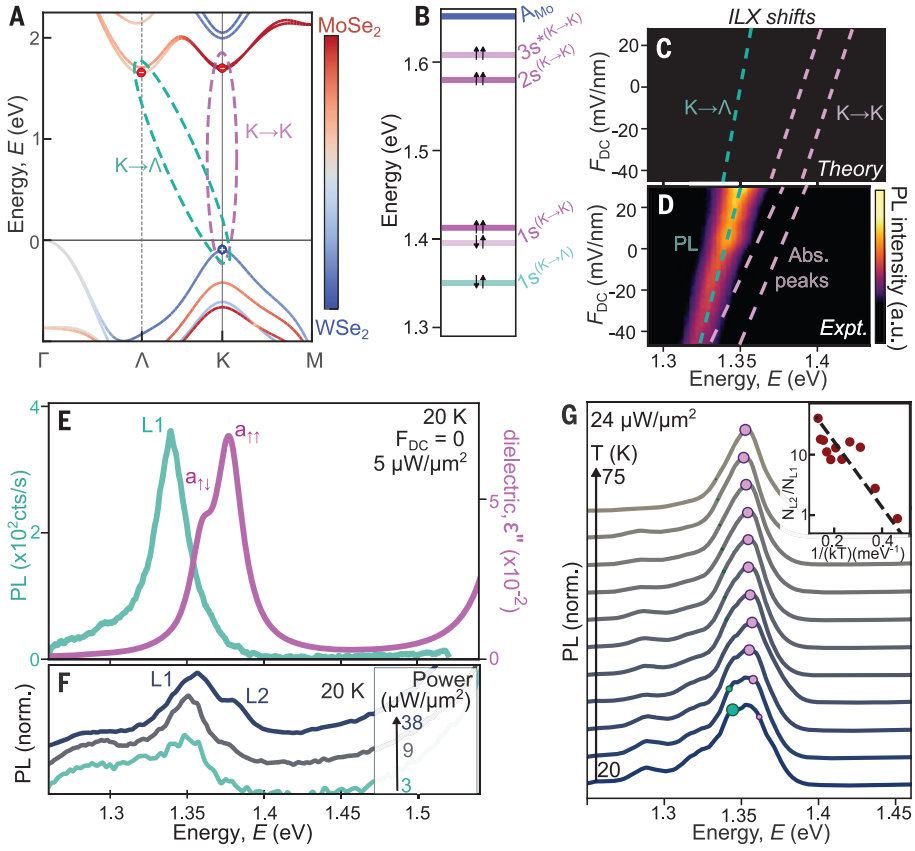


Fig. 3. Origin of ILX photoluminescence. (**A**) Calculated GW band structure of the aligned (H_h^n) heterobilayer, with layer hybridization shown by color. (**B**) Calculated GW-BSE exciton resonances with momentum, spin, and Rydberg assignments. The opacity of the magenta lines ($K \rightarrow K$) indicates relative oscillator strength (table S8). (**C**) Theoretically predicted energies under F_{DC} for the $K \rightarrow \Lambda$ transition ($E_{K\rightarrow\Lambda,0}$ = 1.35 eV, $p_{K\rightarrow\Lambda}$ = 2.5 e·Å), shown in teal, and two lowest-energy, direct $K \rightarrow K$ transitions ($E_{K\rightarrow K,0}$ = 1.396 and 1.413 eV, both with $p_{K\rightarrow K}$ = 4.5 e·Å), in magenta. (**D**) Experimental energy shifts for emission (E_{L1} = 1.340 eV, p_{L1} = 3.48 ± 0.02 e·Å), in teal, and absorption (E_a = 1.359 and 1.377 eV with p_a = 6.2 ± 0.7 and 5.7 ± 0.6 e·Å, respectively), in magenta, under F_{DC} plotted over the PL heatmap (730 nm using 5 μW/μm^2). Note that the dipole ratio of absorption peaks to emission peaks (0.57) matches that of theoretical $K \rightarrow K$ to $K \rightarrow \Lambda$ (0.55). (**E**) Zero-field PL (teal curve) with the dielectric function as measured by R_F (magenta) from (D). (**F**) PL for increasing excitation power. A right-hand shoulder, L2, 18 meV higher than the main peak L1, is apparent above 10 μW/μm^2. (**G**) The high-power PL emission with increasing temperature. The teal and pink circles indicate the energies and relative magnitudes of the L1 ($K \rightarrow \Lambda$) and the L2 ($K \rightarrow K, \uparrow\downarrow$) peaks. The inset shows the relative strengths as a function of inverse temperature; the black-dashed fit assumes a radiative rate difference of γ_{L2}/γ_{L1} = 200 ± 10 and an energy difference of $E_{L2} - E_{K\rightarrow\Lambda} \approx$ 10 ± 2 meV, implying that a phonon of 8 meV is involved in the emission from $K \rightarrow \Lambda$.

ILX transitions compatible with the observed spectral shift (*11*). The broad shoulder in the PL emission at energies below the L1 feature is attributed to defect states or locally strained regions, as it saturates at low powers and disappears at high temperatures. Overall, our results indicate that momentum-indirect transitions dominate the ILX emission in a well-aligned WSe_2/$MoSe_2$ heterostructure, but with an energy only slightly below that of the momentum-direct transition seen in our absorption measurements.

Because the $K \rightarrow \Lambda$ and $K \rightarrow K$ transitions are so close in energy, their relative energies and dominance in PL measurements vary for samples with different strain and twist angle, as suggested by recent work (*24*). This situation is illustrated in the PL and absorption of a second, strained and misaligned sample (fig. S5). The similarity in energy of the indirect and direct ILX transitions also explains the seemingly contradictory claims of $K \rightarrow K$ excitons (*4*, *8*, *10*) and $K \rightarrow \Lambda$ excitons (*9*, *13*, *27*) in the literature. Our measurements allow direct comparison between the $K \rightarrow K$ ILX dipole moments extracted from the absorption spectrum and the ILX PL, thus avoiding the experimental uncertainties encountered when measuring the absolute values of the PL dipole moment [e.g., arising from the hBN dielectric constant (*28*, *29*)].

Apart from providing a comparison for PL measurements, the electromodulation technique introduced here allows the direct determination of the ILX contribution to the dielectric function. The ILX radiative lifetimes thereby established from the measured oscillator strengths are free from the influence of nonradiative processes, while the inferred transition energies are not affected by defects and localized strain that can dominate the emission spectra. The robust values for the ILX radiative lifetimes and energies have allowed us to explore the role of twist angle and moiré potential in the WSe_2/$MoSe_2$ system. Aside from offering fundamental understanding, this knowledge is essential for the potential application of heterobilayer systems in optoelectronic devices that make use of the ready tunability and long lifetime of the ILX. The quantitative characterization of the ILX dielectric response, in combination with the discovery of ILXs at longer wavelengths (*6*), supports the design of systems that integrate these materials into state-of-the-art photonic platforms.

REFERENCES AND NOTES

1. M. Cardona, P. Y. Yu, *Fundamentals of Semiconductors* (Springer, 2010).
2. G. Wang *et al.*, *Rev. Mod. Phys.* **90**, 021001 (2018).
3. J. Pei, J. Yang, T. Yildirim, H. Zhang, Y. Lu, *Adv. Mater.* **31**, e1706945 (2019).
4. P. Rivera *et al.*, *Nat. Nanotechnol.* **13**, 1004–1015 (2018).
5. L. A. Jauregui *et al.*, *Science* **366**, 870–875 (2019).
6. O. Karni *et al.*, *Phys. Rev. Lett.* **123**, 247402 (2019).
7. K. Tran *et al.*, *Nature* **567**, 71–75 (2019).
8. K. L. Seyler *et al.*, *Nature* **567**, 66–70 (2019).
9. A. T. Hanbicki *et al.*, *ACS Nano* **12**, 4719–4726 (2018).
10. L. Zhang *et al.*, *Phys. Rev. B* **100**, 041402 (2019).
11. See supplementary materials.
12. T. Wang *et al.*, *Nano Lett.* **20**, 694–700 (2020).
13. B. Miller *et al.*, *Nano Lett.* **17**, 5229–5237 (2017).
14. C. Robert *et al.*, *Phys. Rev. B* **96**, 155423 (2017).
15. H. Yu, G.-B. Liu, W. Yao, *2D Mater.* **5**, 035021 (2018).
16. J. Choi *et al.*, *Phys. Rev. Lett.* **126**, 047401 (2021).
17. N. Leisgang *et al.*, *Nat. Nanotechnol.* **15**, 901–907 (2020).
18. L. Zhang *et al.*, *Nat. Commun.* **11**, 5888 (2020).
19. M. Goryca *et al.*, *Nat. Commun.* **10**, 4172 (2019).
20. M. K. L. Man *et al.*, *Sci. Adv.* **7**, eabg0192 (2021).
21. H. Yu, Y. Wang, Q. Tong, X. Xu, W. Yao, *Phys. Rev. Lett.* **115**, 187002 (2015).
22. X. Lu, X. Li, L. Yang, *Phys. Rev. B* **100**, 155416 (2019).
23. C. Jin *et al.*, *Nature* **567**, 76–80 (2019).
24. S. Shabani *et al.*, *Nat. Phys.* **17**, 720–725 (2021).
25. P. K. Nayak *et al.*, *ACS Nano* **11**, 4041–4050 (2017).
26. M. H. Naik, S. Kundu, I. Maity, M. Jain, *Phys. Rev. B* **102**, 075413 (2020).
27. R. Gillen, J. Maultzsch, *Phys. Rev. B* **97**, 165306 (2018).
28. L. Wang, Y. Pu, A. K. Soh, Y. Shi, S. Liu, *AIP Adv.* **6**, 125126 (2016).
29. S. K. Jang, J. Youn, Y. J. Song, S. Lee, *Sci. Rep.* **6**, 30449 (2016).

ACKNOWLEDGMENTS

Funding: This research program was supported by the AMOS program, Chemical Sciences, Geosciences, and Biosciences Division, Basic Energy Sciences (BES), Office of Science (SC), US Department of Energy (DOE). Support for sample fabrication was provided at Stanford by the Betty and Gordon Moore Foundation's EPiQS Initiative through grant GBMF9462 and at UC Riverside by the NSF Division of Materials Research through CAREER award 1945660. Sample fabrication made use of the facilities in the Stanford Nano Shared Facilities (SNSF), supported by NSF award ECCS-2026822. WSe_2 crystal growth was supported at Columbia University under the NSF Materials Research Science and Engineering Center through grants DMR-1420634 and DMR-2011738, and hBN crystal growth was

supported by the Elemental Strategy Initiative conducted by MEXT, Japan (grant JPMXP0112101001) and JSPS KAKENHI (grants 19H05790, 20H00354 and 21H05233). Theory was supported by the Center for Computational Study of Excited State Phenomena in Energy Materials (C2SEPEM), funded by DOE-BES, Materials Sciences and Engineering Division, under contract DE-AC02-05CH11231, as part of the Computational Materials Sciences Program. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a DOE-SC User Facility located at Lawrence Berkeley National Laboratory, operated under contract DE-AC02-05CH11231 using NERSC awards BES-ERCAP 0018289 and 0021335. Support for theoretical research at the Weizmann Institute was provided by the Leah Omenn Career Development Chair and Peter and Patricia Gruber Awards. We gratefully acknowledge fellowship support for S.R.-A. (Alon Fellowship), E.B. (Natural Science and Engineering Research Council of Canada under fellowship PGSD3-502559-2017), O.K. (Koret Foundation), and H.B.R. (FAPESP under postdoctoral fellowship 2018/04926-9). **Author contributions:** E.B. and T.F.H. conceived the project. K.W., T.T., B.K., and K.B. supplied raw materials for sample fabrication. E.B., O.K., E.L., A.L.O., and H.B.R. fabricated the samples with assistance from L.Y. E.B. designed the experimental setup. E.B., A.L.O., and X.C. performed the experimental measurements with support from O.K. E.B., O.K., F.H.d.J., and S.R.-A.. interpreted experimental results. F.H.d.J. and S.R.-A. performed theoretical calculations. A.L.O., E.B., H.B.R., L.Y., and O.K. were supervised by T.F.H. E.L. was supervised by C.H.L. All authors contributed to discussions and manuscript preparation. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are available in the manuscript or the supplementary materials.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abm8511
Materials and Methods
Supplementary Text
Figs. S1 to S11
Tables S1 to S8
References (*30*–*56*)

5 November 2021; accepted 23 March 2022
10.1126/science.abm8511

ORGANIC CHEMISTRY

Controlling Ni redox states by dynamic ligand exchange for electroreductive Csp3–Csp2 coupling

Taylor B. Hamby, Matthew J. LaLama, Christo S. Sevov*

Cross-electrophile coupling (XEC) reactions of aryl and alkyl electrophiles are appealing but limited to specific substrate classes. Here, we report electroreductive XEC of previously incompatible electrophiles including tertiary alkyl bromides, aryl chlorides, and aryl/vinyl triflates. Reactions rely on the merger of an electrochemically active complex that selectively reacts with alkyl bromides through 1e$^-$ processes and an electrochemically inactive Ni^0(phosphine) complex that selectively reacts with aryl electrophiles through 2e$^-$ processes. Accessing Ni^0(phosphine) intermediates is critical to the strategy but is often challenging. We uncover a previously unknown pathway for electrochemically generating these key complexes at mild potentials through a choreographed series of ligand-exchange reactions. The mild methodology is applied to the alkylation of a range of substrates including natural products and pharmaceuticals.

Advances in Ni-based catalysis have enabled new C–C bond-forming methodologies that directly couple two C electrophiles in a net-reductive process without the need to preform a nucleophilic coupling partner (*1*–*4*). The ubiquity of simple organohalides or other C electrophiles has caused Ni-catalyzed cross-electrophile coupling (XEC) to become one of the most common strategies for C–C coupling in industry, particularly to form C(sp^3)–C(sp^2) bonds (*5*–*11*). Although chemical (*3*), photoredox (*12*), and electrochemical (*13*) approaches have been developed to deliver the reducing equivalents needed for XEC, all three reductive strategies are limited to couplings of similar classes of alkyl and aryl halides. An example of this substrate-specific reactivity is highlighted in Fig. 1A from our own work on electrochemical XEC (eXEC), in which coupling reactions of aryl bromides and primary (1°) or secondary (2°) alkyl bromides are often quantitative, whereas those of tertiary (3°) alkyl bromides fail to form any cross-products (*14*). This substantial difference in yield is not unique to eXEC because XEC reactions of 3° alkyl bromides are rare (*15*) and the few known examples require aryl iodides or activated (e-deficient) aryl bromides as the coupling partner (*16*–*18*). More broadly, a survey of organohalides that are successfully coupled under any of the three reductive approaches reveals a narrow chemical space of electrophiles that can be paired. XEC of aryl iodide/bromide + 1°/2° alkyl bromide (Fig. 1B, bottom left, dark blue) can be reliably performed in high yield with a wide range of catalysts. By contrast, reactions of substrate combinations that deviate from this constraint, such as those of 3° alkyl bromides or e-rich aryl bromides, are challenging (Fig. 1B, light red and light blue). In addition to the underdeveloped couplings at the boundary of known XEC reactions, more than half of the combinations in Fig. 1B lie beyond the current chemical space for XEC. Namely, couplings of widely available electrophiles such as aryl chlorides or triflates are currently unknown with any alkyl bromide. This work circumvents the limitations of alkyl-aryl XEC. The developed methodology enables couplings of a wide range of unknown or low-yielding combinations of electrophiles (summarized in Fig. 1B, red).

Department of Chemistry and Biochemistry, The Ohio State University, Columbus, OH 43210, USA.
*Corresponding author. Email: sevov.1@osu.edu

Conceptually, cross-product formation in XEC relies on the sequential activation of each electrophile at a low-valent metal complex, most often a Ni complex of pyridyl-based ligands (e.g., 2,2′-bipyridine) (*3*, *19*). Aryl electrophiles are activated by 2e$^-$ oxidative addition at Ni, and alkyl electrophiles react through 1e$^-$ processes to form alkyl radicals (*20*, *21*). Although it initially seemed that activation of each electrophile occurred at a distinct oxidation state of Ni (Ni^0 or Ni^I), a growing body of evidence from electrochemical (*22*), photoredox (*23*), and chemical (*24*) studies suggests that only Ni^I(pyridyl) intermediates are accessible under reductive conditions. On the basis of these reports, we hypothesized that XEC is restricted to electrophiles that react with comparable rates at Ni^I, whereas electrophiles that are exceedingly reactive (3° alkyl halides) or unreactive (Ar–Cl/OTf) at Ni^I are incompatible coupling partners (Fig. 1C) (*20*, *25*, *26*). This competition between activation of alkyl and aryl electrophiles is circumvented in conventional Suzuki or Negishi methodologies that preactivate aryl halides in separate synthetic sequences as organoboron or organozinc reagents, respectively (*27*–*30*). However, even these methodologies remain underdeveloped for reactions of 3° alkyl bromides. Alternatively, alkyl-aryl Suzuki couplings with preformed 3° alkyl boron reagents suffer from similar limitations as XEC reactions, in which only e-deficient aryl electrophiles that readily react with Ni^I are compatible (*31*, *32*).

One potential solution to the limitations pervading both XEC and conventional cross-coupling reactions is to access Ni^0 complexes that could preferentially undergo 2e$^-$ reactions with aryl electrophiles over 1e$^-$ reactions with alkyl bromides. However, electrochemically generating Ni^0(pyridyl) complexes while bypassing the Ni^I state that rapidly reacts with 3° alkyl bromides is challenging. Specifically, (pyridyl)Ni^{II} complexes often exhibit discrete 1e$^-$ redox couples, rather than 2e$^-$ redox couples, and require reduction to Ni^I before a second reduction forms Ni^0 at a more negative potential (Fig. 1C) (*33*, *34*). Even when Ni^0 is formed, the complex undergoes rapid comproportionation with remaining Ni^{II} in solution

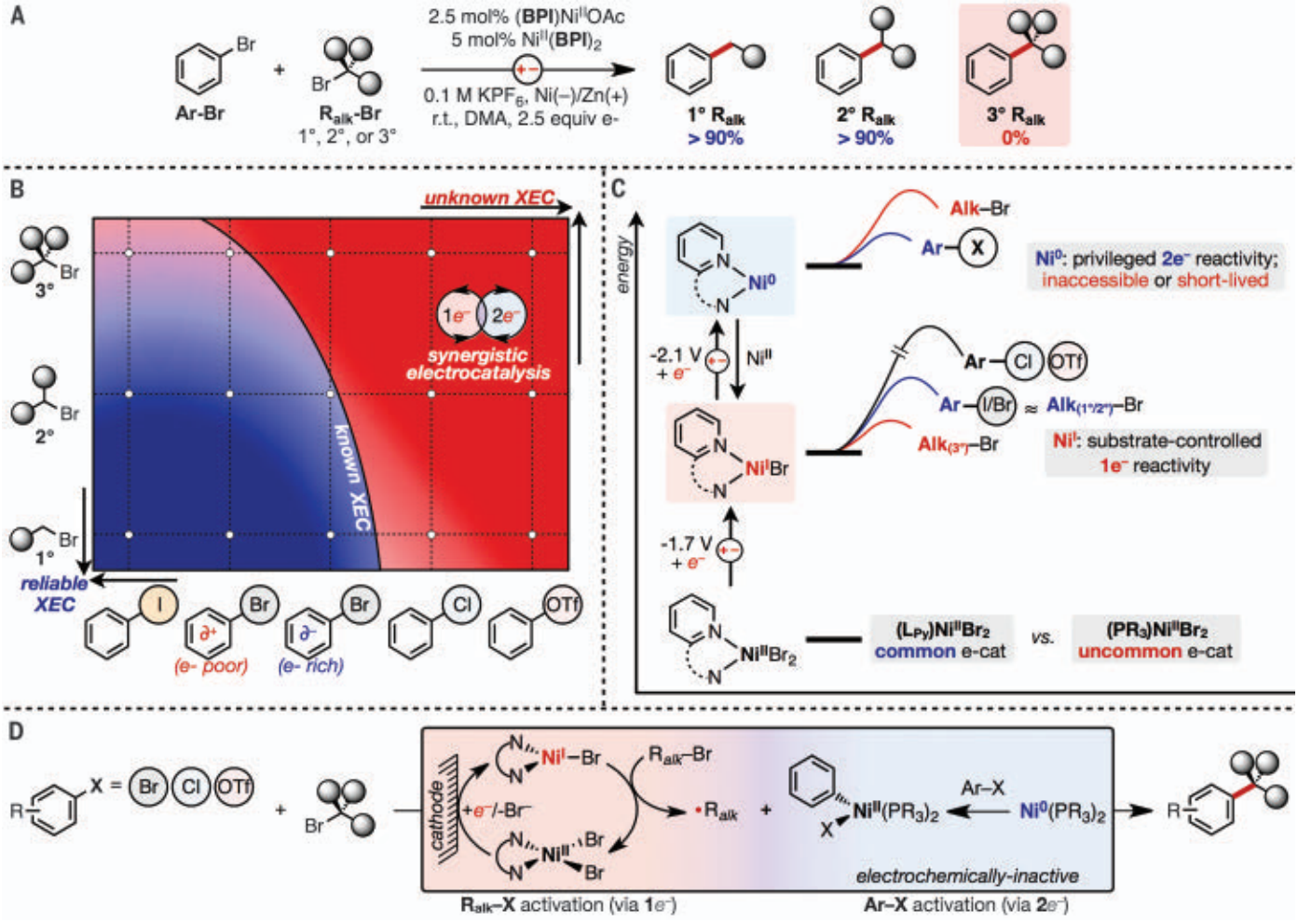


Fig. 1. Background, limitations, and design for C(sp^2)-C(sp^3) XEC. (**A**) Example of substrate-controlled limitations in XEC. DMA, dimethylacetamide; r.t., room temperature; BPI, bis(2-pyridylimino)isoindoline. (**B**) Reported outcomes of XEC reactions of varying classes of alkyl and aryl electrophiles. Blue indicates high-yielding/reliable combinations; red, low-yielding/unknown combinations. (**C**) Qualitative energy diagram illustrating the reductive accessibility of Ni^I or Ni^0 complexes and their relative reactivities with various aryl or alkyl electrophiles. Reported $E_{1/2}$ values (*14*) are for the electrocatalyst in (A) and referenced to ferrocene, Fc/Fc^+. (**D**) Proposed strategy for XEC that decouples substrate activation to catalysts with dedicated $1e^-$ or $2e^-$ reactivity.

to form the undesired Ni^I intermediate (*22*, *35*, *36*). By contrast, Ni(phosphine) complexes are known to react through $Ni^{0/II}$ processes with a wide range of aryl electrophiles, including aryl chlorides and ethers (*37*–*40*). Schoenebeck's group recently contrasted the reactivities of phosphine- versus bipyridine-ligated Ni complexes for C–S coupling reactions, noting that Ni^0(phosphine) complexes are accessible and catalyze reactions of aryl chlorides. Conversely, reactions catalyzed by complexes of pyridyl analogs react through the Ni^I state and are limited to couplings of aryl iodides or bromides (*41*). Despite the prevalence of phosphine-based complexes in Ni-catalyzed Suzuki- and Negishi-type catalysis, they are rarely used in electrosynthesis or even photoredox catalysis (Fig. 1C, right) (*13*, *42*, *43*). Electroreduction of neutral Ni(phosphine) complexes in polar solvents can be challenging (see mechanistic investigation below), necessitating stabilizers (e.g., hexamethylphosphoramide) or redox promoters (*44*, *45*). Moreover, the rare examples of XEC catalyzed by phosphine-ligated complexes require Grignard reagents as reductants (*46*), which further highlights the challenge of reductively activating non-pyridyl Ni complexes. We viewed the poor electrochemical activity of Ni(phosphine) complexes under certain conditions as an opportunity to evade $1e^-$ electrochemical events at Ni and possibly promote a $Ni^{0/II}$ manifold that preferentially activates aryl over alkyl electrophiles. Furthermore, an electrochemically active (pyridyl)Ni complex would generate alkyl radicals through complementary $1e^-$ reactions upon electroreduction. Activation of one electrophile is thereby decoupled from activation of the other (Fig. 1D).

Our initial studies sought to establish the feasibility of a dual-catalyst approach by targeting XEC reactions of 3° alkyl bromides. Pseudo-stoichiometric reactions were performed at high Ni loadings (30 mol%) in the one-pot, two-step sequence illustrated in Fig. 2. First, $Ni(COD)_2$ and a phosphine were combined with a mixture of 4-butylbromobenzene and *tert*-butyl bromide with the aim of selectively forming the Ni^{II}(aryl) intermediate. The mixture was then electrolyzed in the presence of (bpp)$NiBr_2$ (**1**, bpp = 2,6-bispyrazolylpyridine), which was identified as an electrocatalyst that rapidly reacts with 3° alkyl bromides upon reduction (see the supplementary materials, figs. S6 and S11). Yields <30% from this assay would indicate only stoichiometric reactivity, and those >30% would indicate turnover of the Ni complex. Although reactions performed with most of the tested phosphines formed the target product in yields <30%, promising yields from reactions with PHOX and Quinap led us to evaluate ligands with a similar architecture. In particular, reactions

with the isopropyl analog of quinazolinap (**iPrQ**) generated products in 61% yield, indicating that catalytic turnover was possible.

The combination of phosphine **iPrQ** and radical generator **1** was further developed for catalytic eXEC, as summarized in Table 1. Reactions were performed under constant-current electrolysis at room temperature with a Ni-foam cathode and a Zn anode in DMF/KPF_6 as electrolyte. We discovered that reactions performed with the Ni complex of bpp (**1**) were high yielding for the coupling of an electron-rich aryl bromide and *tert*-butyl bromide to form product ***5*** but also formed a small yet problematic quantity of the inseparable byproduct ***5-iso*** (entry 2). Such isomeric byproducts are common in the rare examples of XEC reactions of 3° alkyl electrophiles after alkyl isomerization at Ni through β-hydride elimination and reinsertion to form a stable primary alkyl-Ni complex (*15*). By contrast, reactions performed with a combination of (bpp)$MnCl_2$, $NiCl_2$•dme, and **iPrQ** resulted in the exclusive formation of product ***5*** in 75% yield (entry 1). Cyclic voltammetry (CV) studies revealed that reduced (bpp)$MnCl_2$ does not react with alkyl bromides and that Mn is readily displaced by Ni^{II} to form complex **1** in situ (see the supplementary materials, figs. S4, S5, and S18). Although the Ni analog **1** is responsible for alkyl radical formation upon electroreduction, the low concentration of remaining (bpp)$MnCl_2$ may serve as an additional source of bpp ligand to promote the formation of key catalytic intermediates that mediate C–C coupling of alkyl radicals without isomerization (see below).

Other redox-active complexes that could activate 3° alkyl bromides were similarly evaluated, but reactions formed ***5*** in low yields and with poor selectivity over ***5-iso*** (entries 3 and 4). Entries 5 to 8 summarize results from reactions performed at varying currents or constant potentials at the observed cell voltage of the standard conditions ($E_{anode\text{-}cathode}$ = 0.4 to 0.6 V). Control experiments confirmed that Ni, (bpp)$MnCl_2$, and phosphine are all necessary for product formation, and the omission of any one component resulted in only trace conversion of the aryl bromide (entries 9 to 11). However, electrolysis is not essential to the reaction. Reactions performed with a Zn anode but without an applied potential formed ***5*** in 17% yield (entry 12), and higher yields of 60% could be obtained from reactions with an excess of activated Zn powder after 24 hours (entry 13). Nonetheless, these reactions formed substantial quantities of ***5-iso***, which highlights the importance of electrochemically controlling the rate of catalyst activation through applied current. Finally, reactions were highest yielding when performed with an equimolar loading of **iPrQ** to Ni (entries 14 and 15) and a slight excess of aryl bromide (1.5 equivalents). Reactions conducted with a 1:1 ratio of aryl to alkyl electrophile resulted in yields of only 57% (entry 16). This low yield stems from the complete consumption of the aryl bromide rather than the 3° alkyl bromide. The relative reactivity of the electrophiles under these conditions contrasts conventional XEC reactions, in which the alkyl halide is preferentially consumed over the aryl halide.

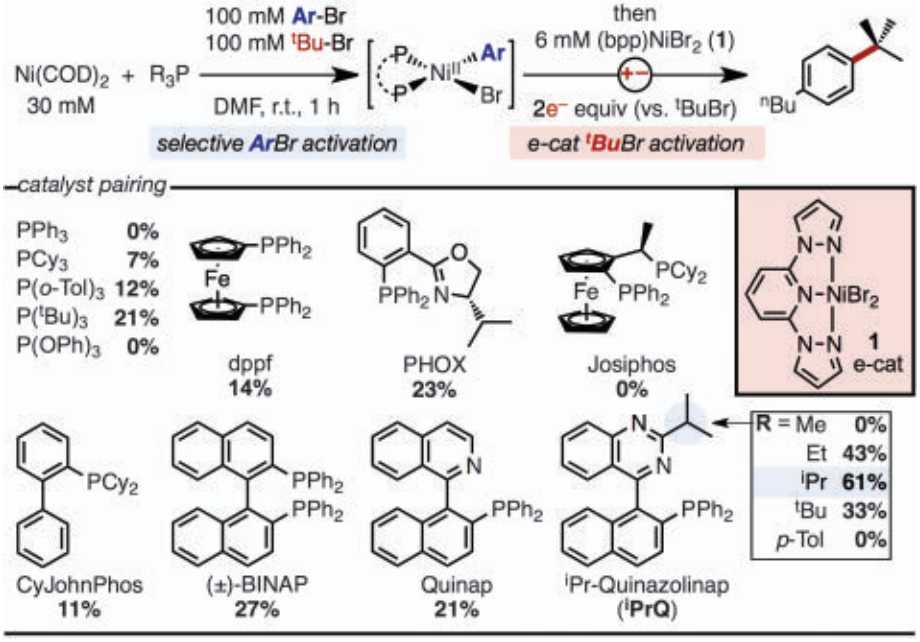
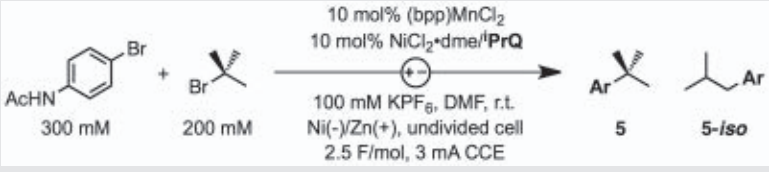


Fig. 2. Identification of catalyst combinations for dual XEC. Evaluation of Ni-phosphine combinations paired with electrocatalyst **1** for XEC. Calibrated gas chromatography (GC) yields based on ^{t}BuBr are reported. DMF, dimethylformamide COD, cyclooctadiene.

Table 1. Development of catalytic reaction conditions.



Entry	Deviation from standard conditions	%conv ArBr	%yield 5	%yield 5-*iso*
1	Standard conditions	86	75	0
2	(bpp)$NiCl_2$ instead of [(bpp)$MnCl_2$ + $NiCl_2$•dme]	92	83	8
3	(bpp)$Co(OAc)_2$ instead of (bpp)$MnCl_2$	72	35	6
4	(bpp)$FeBr_2$ instead of (bpp)$MnCl_2$	74	29	11
5	Constant current: 6 mA	82	49	1
6	Constant current: 1.5 mA	62	45	0
7	Constant voltage: E_{cell} = 0.4 V	81	47	0
8	Constant voltage: E_{cell} = 0.6 V	89	53	1
9	No $NiCl_2$•dme	0	0	0
10	No (bp)$MnCl_2$	7	0	0
11	No **iPrQ** ligand	6	5	1
12	Zn^0 anode, no e-chem	24	17	1
13	Zn^0 powder (3 equiv), no e-chem	74	60	5
14	5 mol% **iPrQ**	78	43	1
15	15 mol% **iPrQ**	84	68	0
16	200 mM ArBr and 200 mM ^{t}BuBr	100	57	0

Calibrated GC yields against an internal standard are reported. CCE, constant current electrolysis.

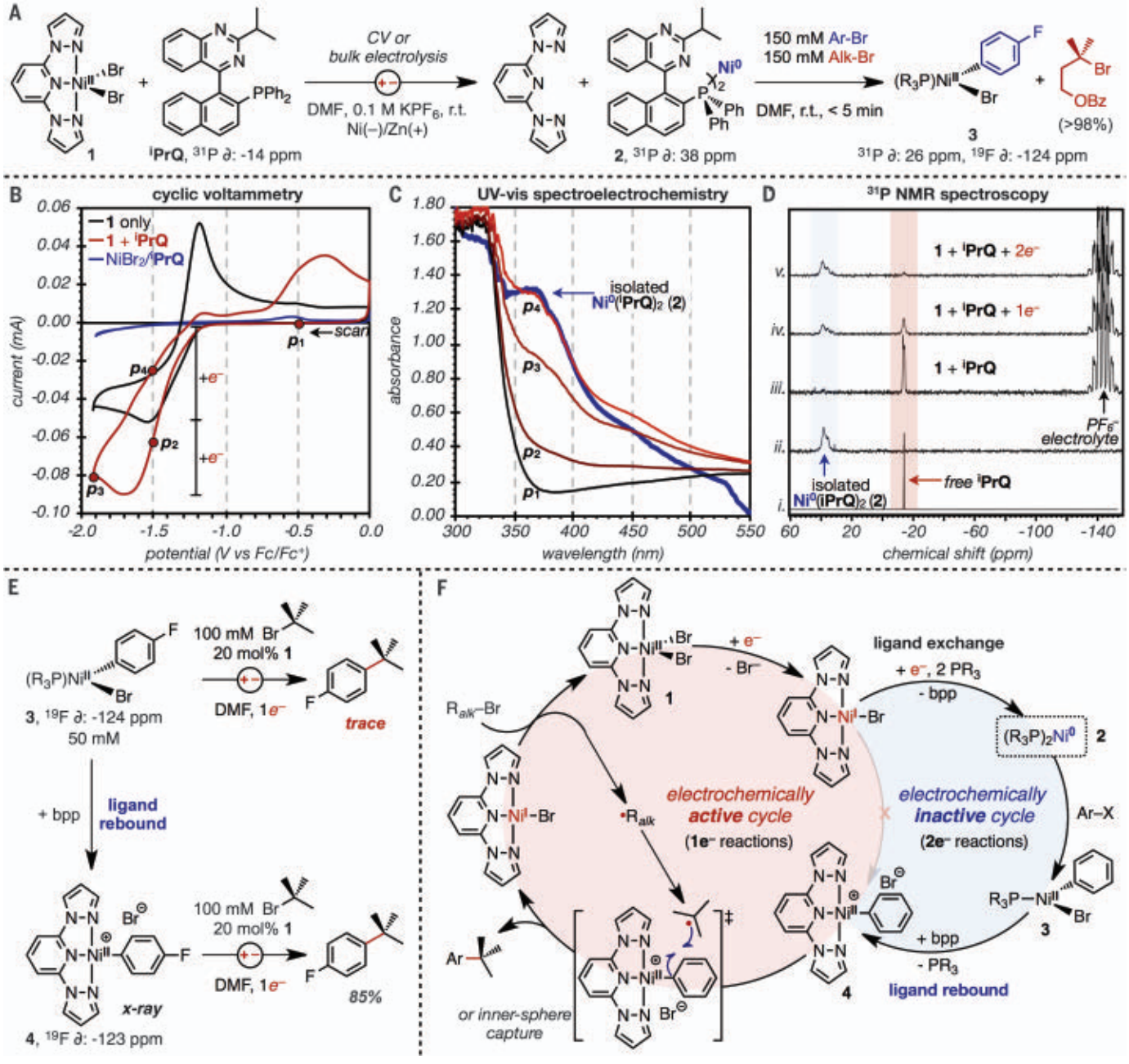


Fig. 3. Mechanistic investigation. (**A**) Conditions for electrolysis of **1** in the presence of **iPrQ** to form Ni^0(**iPrQ**)$_2$ (left). Subsequent reactions of isolated **2** with an alkyl or aryl electrophile to form the aryl complex **3** exclusively (right). (**B**) CVs of **1** (10 mM) with and without **iPrQ** (10 mM). Conditions: 0.1 M KPF_6/DMF, glassy C WE, Pt CE, 100 mV/s. (**C**) Spectroelectrochemical analysis from CVs in (B). Only four points from the CV are plotted for clarity along with the UV-vis spectrum of **2**. (**D**) Conversion of **1** to **2** during bulk electrolysis as monitored by ^{31}P NMR spectroscopy. (**E**) Electrolysis reactions of isolated Ni(aryl) intermediates **3** and **4** with *tert*-butyl bromide. (**F**) Proposed mechanism.

We next investigated the mechanism of the reaction (Fig. 3) before evaluating the full scope of the methodology because it was unclear how aryl activation could occur beyond the stoichiometric approach outlined in Fig. 2. Control experiments and CV studies revealed that (bpp)$NiBr_2$ **1** is reduced at a mild potentials ($E_{1/2}$ = –1.4 V versus Fc/Fc^+; Fig. 3B, black trace) compared with other XEC electrocatalysts (*47*), but the resulting complex was only reactive toward alkyl bromides (see the supplementary materials, fig. S6). By contrast, aryl bromides are readily activated by the combination of **iPrQ** and a Ni^0 precursor, but such low-valent phosphine complexes are not accessible by direct electroreduction. As demonstrated by the blue trace in Fig. 3B, no significant redox event was detected above –2 V from CVs of **iPrQ** and a Ni^{II} salt, highlighting the electrochemical inactivity of this metal-ligand combination under the reaction conditions. However, we noted an unusual effect of added **iPrQ** on the electroreduction of complex **1**. CVs of this mixture exhibited the same onset potential for reduction as **1** alone, but the amplitude of the reductive current nearly doubled (red versus black traces). This result suggests that 2e$^-$, rather than 1e$^-$, are transferred at the reduction potential of **1** when phosphine is added. We performed spectroelectrochemical analysis to gain insight into the intermediates formed during reduction. As illustrated in Fig. 3C, the ultraviolet-visible

(UV-vis) spectra acquired at points **p1** to **p4** during the CV scan revealed the growth of an absorbance at 370 nm. The UV-vis spectrum acquired at **p4** almost perfectly matched that of the Ni^0(**iPrQ**) bisphosphine (**2**) that we isolated and characterized from a reaction of $Ni(COD)_2$ and **iPrQ**.

Surprised that a Ni^0(phosphine) complex might be generated at such mild potentials, we conducted bulk electrolysis of complex **1** in the presence of **iPrQ** and analyzed solution aliquots by ^{31}P nuclear magnetic resonance (NMR) spectroscopy throughout the reaction (Fig. 3D). As a reference, free **iPrQ** exhibited a ^{31}P resonance at –14 ppm, whereas the ligated

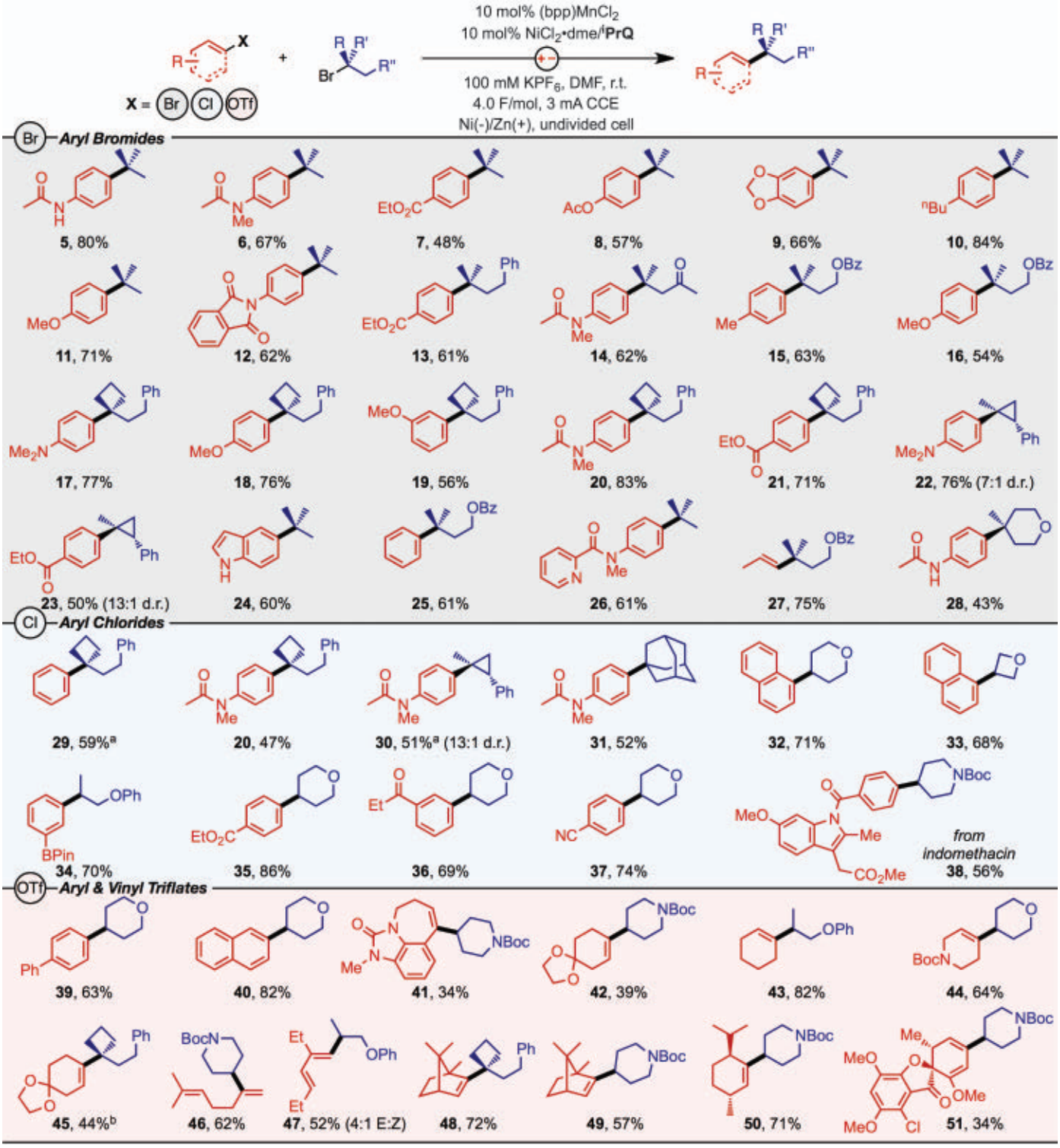


Fig. 4. Reaction scope. Isolated yields from XEC reactions are shown, organized by C(sp^2) electrophile (Br, Cl, and OTf). Conditions: 200 mM alkyl bromide, 300 to 400 mM aryl electrophile, 2.5 ml solution, reaction temperature ranging from room temperature to 60°C, reaction time ranging from 12 to 18 hours. See the supplementary materials for specific reaction conditions. [a]Calibrated GC yield. [b]NMR yield.

complex **2** exhibited a broad resonance at 38 ppm (spectra *i* and *ii*, respectively). Under the electrochemical conditions, the added **iPrQ** remained uncoordinated, indicating that a bpp-ligated Ni^{II} is more stable than a **iPrQ**-ligated Ni^{II} (spectrum *iii*). However, spectra acquired after reduction by $1e^-$ and $2e^-$ equivalents relative to **1** revealed a conversion of free **iPrQ** to complex **2** (spectra *iv* and *v*). These data provide further support that the Ni^0(phosphine) complex is formed at mild potentials and that it persists in the bulk solution without undergoing comproportionation with remaining Ni^{II} complexes to form Ni^I. Most importantly, stoichiometric reactions of **2** with a combination of *para*-fluorobromobenzene and a 3° alkyl bromide revealed exclusive activation of the aryl bromide (>95% conversion versus Ni), whereas none of the 3° alkyl bromide reacted (Fig. 3A, right). The product from this reaction was characterized as ***3*** by ^{19}F and ^{31}P NMR spectroscopy (*48*). These studies underscore that the accessibility and persistence of a Ni^0 intermediate is a critical feature that enables preferential activation of aryl over alkyl electrophiles.

Finally, we investigated the reactivity of Ni^{II}(aryl) ***3*** toward radical capture and C–C coupling with 3° alkyl bromides. Electroreduction of a solution containing ***3***, *tert*-butyl bromide, and **1** as the radical generator with $1e^-$ equivalent (versus ***3***) formed only trace quantities of coupled product (Fig. 3E). This unexpected result indicates that the phosphine complex ***3*** is incompetent toward radical capture and product formation. We realized that electroreductive formation of the phosphine complexes in catalytic reactions releases one equivalent of the bpp ligand, whereas the stoichiometric studies from ***3*** bypass those steps and lack any bpp. Therefore, we added one equivalent of bpp to complex ***3*** and observed the immediate displacement of **iPrQ** in ^{31}P NMR spectra along with a new ^{19}F resonance at –123 ppm. The resulting species was isolated in 68% yield and characterized as the [(bpp)Ni^{II}(aryl)]Br complex **4** by x-ray diffraction (refer to fig. S22 for ORTEP). In contrast to C–C coupling attempts from ***3***, the analogous experiment from **4** formed the coupled product in 85% yield. Phosphine-bpp exchange to form **4** similarly occured when (bpp)$MnCl_2$ was added to ***3***. The weakly bound Mn-bpp complex likely serves as a reservoir for bpp to promote the second ligand-exchange reaction before radical capture (see the supplementary materials, figs. S18 and S19).

Collectively, these data are summarized as the proposed catalytic cycle in Fig. 3F. The reaction is initiated by electroreduction of (bpp)$NiBr_2$ (**1**), followed by rapid ligand exchange with concomitant reduction to form a stable Ni^0(phosphine) (**2**). This phosphine complex is electrochemically inactive at the operating potentials ($E_{cathode}$ = –1.5 V) and does not comproportionate to form Ni^I intermediates as is common for Ni(pyridyl) complexes. Rather, **2** undergoes rapid $2e^-$ oxidative addition with aryl electrophiles to form ***3*** even in the presence of highly activated alkyl bromides. The bpp ligand that was displaced upon electroreductive activation or that was weakly bound to Mn^{II} subsequently recoordinates the aryl complex ***3*** in a second ligand exchange reaction to form **4**. This ligand-rebound event is a critical step because product formation only occurs from the bpp-ligated aryl complex **4**. Overall, these studies reveal a complex series of ligand exchange reactions that facilitate electrochemical generation of highly reactive Ni^0(phosphine) complexes, ultimately enabling XEC of 3° alkyl bromides. Through these exchange reactions, the XEC reaction is decoupled into one regime in which electrochemically active complexes undergo $1e^-$ reactions with alkyl electrophiles (bpp complexes **1** and **4**; Fig. 3F, red) and a second regime in which electrochemically inactive complexes preferentially undergo $2e^-$ reactions with aryl electrophiles (phosphine complexes **2** and ***3***; Fig. 3F, blue).

These mechanistic insights suggested that XEC could in principle be performed with any aryl electrophile prone to activation by a Ni^0 (phosphine). Therefore, we targeted reactions of widely available, but previously incompatible, electrophiles: 3° alkyl bromides, aryl chlorides, and aryl/vinyl triflates (Fig. 4). Products from reactions of *tert*-butyl bromide and a range of e-rich and e-deficient aryl bromides at room temperature were formed in high analytical yield (generally 70 to 85%) and were isolated in good yields as single constitutional isomers (**5** to **11**). The mild conditions of the electrocatalytic reaction are critically important because the 3° electrophiles are prone to elimination at elevated temperatures or protodehalogenation under highly reducing conditions. Reactions of more complex alkyl electrophiles that are susceptible to isomerization formed the desired products in good yields (**13** to **16**) with selectivities over isomers that exceeded 30:1. Additionally, alkyl bromides containing strained rings underwent XEC in high yield to form cyclobutanes or cyclopropanes with all-C quaternary centers (**17** to **23**). The mild methodology was also compatible with unprotected indoles (**24**), pyridyl-substituted acetamides (**26**), vinyl bromides (**27**), and substituted tetrahydropyrans (**28**). Overall, we found coupling reactions of 3° alkyl bromides and aryl bromides to be robust and highly reliable.

We next targeted XEC reactions of aryl chlorides. Products from reactions of 3° alkyl bromides and aryl chlorides were all isolated in good yields (**29** to **31**, **20**). We observed parallel conversion of both the aryl chloride and alkyl bromide throughout the reaction, rather than preferential consumption of just the alkyl bromide. Because XEC reactions of aryl chlorides are extremely rare and limited to couplings of unreactive alkyl electrophiles (*49*, *50*), we also included coupling reactions with 2° alkyl bromides (**32** to **38**). A few noteworthy reactions from this combination include the chemoselective coupling at chloride to form the boryl-substituted product (**34**) in high yield and the modification of the chloroaryl fragment of indomethacin to install a BOC-protected piperidine (**38**).

Finally, we extended this methodology to reactions of aryl and vinyl triflates. Although reactions of aryl triflates formed products in high yield (**39** and **40**), we primarily focused our efforts toward developing reactions with vinyl triflates. These electrophilic substrates are attractive coupling partners because they can be prepared in a single step from a wide range of naturally occurring or pharmaceutically relevant ketones and aldehydes. As an example, the vinyl triflate of the zilpaterol precursor was prepared in a single step and directly coupled with an alkyl bromide to form the piperidyl-modified urea **41**. Other vinyl triflates derived from cyclic ketones underwent coupling with a range of alkyl bromides (**42** to **44**), including 3° analogs (**45** and **48**). Reactions with vinyl triflates derived from acyclic ketones (**46**) and aldehydes (**47**) enabled the direct synthesis of highly substituted alkenes and dienes. These olefinic products could be subsequently hydrogenated to form products of a formal alkyl-alkyl coupling reaction (see the supplementary materials, fig. S20). Reactions with vinyl triflates of highly functionalized natural products could also be performed to generate C–C coupled products from a range of alkyl bromides (**48** to **51**).

Overall, the generality of this methodology highlights how an unusual mechanism involving dynamic ligand exchange can be used to control the redox states of Ni and ultimately expand C–C bond-forming methodologies as a whole. XEC reactions of organohalides are thus no longer limited to the canonical couplings of 1°/2° alkyl bromides and aryl iodides/bromides. Beyond the advances to XEC, these findings represent a fundamental solution to reductively accessing highly reactive Ni^0 complexes in preference to the Ni^I intermediates that currently dominate Ni-redox catalysis.

REFERENCES AND NOTES

1. A. Conan, S. Sibille, E. D'Incan, J. Périchon, *J. Chem. Soc. Chem. Commun.* **1**, 48–49 (1990).
2. X. Wang, Y. Dai, H. Gong, *Top. Curr. Chem. (Cham)* **374**, 43 (2016).
3. D. J. Weix, *Acc. Chem. Res.* **48**, 1767–1775 (2015).
4. S. J. Harwood *et al.*, *Science* **375**, 745–752 (2022).
5. S. K. Nimmagadda *et al.*, *Org. Process Res. Dev.* **24**, 1141–1148 (2020).
6. C. C. Le *et al.*, *ACS Cent. Sci.* **3**, 647–653 (2017).
7. R. Duvadie *et al.*, *Org. Process Res. Dev.* **25**, 2323–2330 (2021).
8. R. J. Perkins, A. J. Hughes, D. J. Weix, E. C. Hansen, *Org. Process Res. Dev.* **23**, 1746–1751 (2019).

9. G. L. Beutner *et al.*, *J. Org. Chem.* **86**, 10380–10396 (2021).
10. I. Grimm *et al.*, *Org. Process Res. Dev.* **24**, 1185–1193 (2020).
11. A. W. Dombrowski *et al.*, *ACS Med. Chem. Lett.* **11**, 597–604 (2020).
12. P. Zhang, C. C. Le, D. W. C. MacMillan, *J. Am. Chem. Soc.* **138**, 8084–8087 (2016).
13. C. A. Malapit *et al.*, *Chem. Rev.* **122**, 3180–3218 (2022).
14. B. L. Truesdell, T. B. Hamby, C. S. Sevov, *J. Am. Chem. Soc.* **142**, 5884–5893 (2020).
15. W. Xue *et al.*, *Chem. Soc. Rev.* **50**, 4162–4184 (2021).
16. X. Wang, S. Wang, W. Xue, H. Gong, *J. Am. Chem. Soc.* **137**, 11562–11565 (2015).
17. X. Wang *et al.*, *J. Am. Chem. Soc.* **140**, 14490–14497 (2018).
18. J. Liu, Y. Ye, J. L. Sessler, H. Gong, *Acc. Chem. Res.* **53**, 1833–1845 (2020).
19. J. B. Diccianni, T. Diao, *Trends Chem.* **1**, 830–844 (2019).
20. C. L. Wagner, G. Herrera, Q. Lin, C. T. Hu, T. Diao, *J. Am. Chem. Soc.* **143**, 5295–5300 (2021).
21. J. B. Diccianni, J. Katigbak, C. Hu, T. Diao, *J. Am. Chem. Soc.* **141**, 1788–1796 (2019).
22. Y. Kawamata *et al.*, *J. Am. Chem. Soc.* **141**, 6392–6402 (2019).
23. N. A. Till, S. Oh, D. W. C. MacMillan, M. J. Bird, *J. Am. Chem. Soc.* **143**, 9332–9337 (2021).
24. R. Sun, Y. Qin, D. G. Nocera, *Angew. Chem. Int. Ed.* **59**, 9527–9533 (2020).
25. Q. Lin, Y. Fu, P. Liu, T. Diao, *J. Am. Chem. Soc.* **143**, 14196–14206 (2021).
26. S. Bajo, G. Laidlaw, A. R. Kennedy, S. Sproules, D. J. Nelson, *Organometallics* **36**, 1662–1672 (2017).
27. J. Xu, O. P. Bercher, M. P. Watson, *J. Am. Chem. Soc.* **143**, 8608–8613 (2021).
28. S. L. Zultanski, G. C. Fu, *J. Am. Chem. Soc.* **135**, 624–627 (2013).
29. T. G. Chen *et al.*, *Angew. Chem. Int. Ed.* **58**, 2454–2458 (2019).
30. Z. Wang, Z. P. Yang, G. C. Fu, *Nat. Chem.* **13**, 236–242 2021).
31. D. N. Primer, G. A. Molander, *J. Am. Chem. Soc.* **139**, 9847–9850 (2017).
32. M. Yuan, Z. Song, S. O. Badir, G. A. Molander, O. Gutierrez, *J. Am. Chem. Soc.* **142**, 7225–7234 (2020).
33. T. C. Palmer *et al.*, *ChemSusChem* **14**, 1214–1228 (2021).
34. Q. Lin, G. Dawson, T. Diao, *Synlett* **32**, 1606–1620 (2021).
35. C. Sandford *et al.*, *Chem. Sci.* **10**, 6404–6422 (2019).
36. A. G. Lappin, A. McAuley, *Adv. Inorg. Chem.* **32**, 241–295 (1988).
37. B. L. H. Taylor, E. C. Swift, J. D. Waetzig, E. R. Jarvo, *J. Am. Chem. Soc.* **133**, 389–391 (2011).
38. S. Ge, R. A. Green, J. F. Hartwig, *J. Am. Chem. Soc.* **136**, 1617–1627 (2014).
39. R. J. Somerville, L. V. A. Hale, E. Gómez-Bengoa, J. Burés, R. Martin, *J. Am. Chem. Soc.* **140**, 8771–8780 (2018).
40. B. M. Rosen *et al.*, *Chem. Rev.* **111**, 1346–1416 (2011).
41. I. Kalvet, Q. Guo, G. J. Tizzard, F. Schoenebeck, *ACS Catal.* **7**, 2126–2132 (2017).
42. M. Yan, Y. Kawamata, P. S. Baran, *Chem. Rev.* **117**, 13230–13319 (2017).
43. J. Twilton *et al.*, *Nat. Rev. Chem.* **1**, 0052 (2017).
44. A. Jutand, *Chem. Rev.* **108**, 2300–2347 (2008).
45. B. R. Walker, C. S. Sevov, *ACS Catal.* **9**, 7197–7203 (2019).
46. E. L. Lucas, E. R. Jarvo, *Nat. Rev. Chem.* **1**, 0065 (2017).
47. J. L. S. Zackasee, S. Al Zubaydi, B. L. Truesdell, C. S. Sevov, *ACS Catal.* **12**, 1161–1166 (2022).
48. R. Kehoe *et al.*, *Organometallics* **37**, 2450–2467 (2018).
49. H. A. Sakai, W. Liu, C. C. Le, D. W. C. MacMillan, *J. Am. Chem. Soc.* **142**, 11691–11697 (2020).
50. S. Kim, M. J. Goldfogel, M. M. Gilbert, D. J. Weix, *J. Am. Chem. Soc.* **142**, 9902–9907 (2020).

ACKNOWLEDGMENTS

Funding: This work was supported by the National Institutes of Health (grant R35 GM138373). **Author contributions:** T.B.H. and C.S.S. conceived the work and designed the experiments. T.B.H. and M.J.L. performed all experiments and collected all data. T.B.H. and C.S.S. analyzed the data and wrote the manuscript. M.J.L. provided revisions. **Competing interests:** C.S.S. and T.B.H. are co-inventors on the application filed by The Ohio State University for US patent application number 63/308,319 related to this work. **Data and materials availability:** All experimental data, analytical procedures, and copies of spectra are available in the supplementary materials.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abo0039
Materials and Methods
Figs. S1 to S22
References (*51*–*74*)

6 January 2022; accepted 19 March 2022
10.1126/science.abo0039

SOLAR CELLS

Organometallic-functionalized interfaces for highly efficient inverted perovskite solar cells

Zhen Li[1]†, Bo Li[1]†, Xin Wu[1]†, Stephanie A. Sheppard[2], Shoufeng Zhang[1], Danpeng Gao[1], Nicholas J. Long[2]*, Zonglong Zhu[1,3]*

Further enhancing the performance and stability of inverted perovskite solar cells (PSCs) is crucial for their commercialization. We report that the functionalization of multication and halide perovskite interfaces with an organometallic compound, ferrocenyl-bis-thiophene-2-carboxylate ($FcTc_2$), simultaneously enhanced the efficiency and stability of inverted PSCs. The resultant devices achieved a power conversion efficiency of 25.0% and maintained >98% of their initial efficiency after continuously operating at the maximum power point for 1500 hours under simulated AM1.5 illumination. Moreover, the $FcTc_2$-functionalized devices passed the international standards for mature photovoltaics (IEC61215:2016) and have exhibited high stability under the damp heat test (85°C and 85% relative humidity).

Power conversion efficiencies (PCEs) as high as 25.7% have been realized for single-junction conventional n-i-p perovskite solar cells (PSCs), approaching the PCEs of state-of-the-art crystalline-silicon solar cells (*1*–*3*). Inverted (p-i-n structure) devices, with a deposition sequence of hole-transport (p), intrinsic (i), and electron-transport (n) layers, have exhibited greater stabilities and lifetimes because of their undoped hole-transporting layers (HTLs) and the formation of highly crystalline perovskite films (*4*–*10*). Recently, strategies for managing defects and ion migration in inverted PSCs have further contributed to device stability. For example, Chen *et al.* have used solid-state carbohydrazide to modulate the crystallization of perovskites and fabricate a minimodule that maintained 85% of its initial PCE under 1-sun illumination (where 1 sun is defined as the standard illumination at AM1.5, or 1 kW m^{-2}) for 1000 hours (*11*), and Bai *et al.* have applied ionic liquids within perovskite films by suppressing ion migration and have fabricated PSCs that exhibited only ~5% degradation of device performance under continuous simulated AM1.5 irradiation for >1800 hours (*12*). However, the currently reported operational lifetime of inverted PSCs under international standards still lags far behind the 25-year lifetime guarantee for commercialized silicon solar cells (*2*, *13*). Moreover, although synergistic tailoring of grains, defects, and interfaces can boost efficiencies to 23.3%, there is still a lack of strategy that could result in efficiencies of up to 25% to rival n-i-p PSCs and silicon solar cells (*14*–*16*).

[1]Department of Chemistry, City University of Hong Kong, Kowloon 999077, Hong Kong. [2]Department of Chemistry, Imperial College London, White City Campus, London, UK. [3]Hong Kong Institute of Clean Energy, City University of Hong Kong, Kowloon 999077, Hong Kong.
***Corresponding author. Email: zonglzhu@cityu.edu.hk (Z.Z.); n.long@imperial.ac.uk (N.J.L.)**
†These authors contributed equally to this work.

Here, we report highly efficient and stable inverted PSCs through interface functionalization with an organometallic compound, ferrocenyl-bis-thiophene-2-carboxylate ($FcTc_2$), that not only provided strong chemical Pb-O binding to reduce surface trap states but also accelerated interfacial electron transfer through the electron-rich and electron-delocalizable ferrocene units. The improved interface binding and carrier transport properties of $FcTc_2$ contribute to superior device stability. The resulting devices achieved a PCE of 25.0% (with certified 24.3%) and maintained >98% of their initial efficiency in long-term operational stability tests with continuous 1-sun illumination for >1500 hours. Moreover, the $FcTc_2$-treated devices exhibited excellent stability under damp heat tests [85°C and 85% relative humidity (RH)], which have passed the international standards for silicon solar cells.

Interfaces were functionalized with $FcTc_2$ by using the inherent carboxylate and thiophene groups around the central ferrocene motif (see Fig. 1A for chemical structure). The ultraviolet-visible (UV-vis) absorption spectra of $FcTc_2$ in solution and as a thin film are shown in figs. S1 and S2, respectively. The device configuration is depicted in Fig. 1A, in which poly(triaryl amine) (PTAA) was the HTL and C_{60} was the electron transfer layer (ETL). Figure S3 shows the cross-sectional scanning electron microscopy (SEM) image of a typical device with a perovskite composition of $Cs_{0.05}(FA_{0.98}MA_{0.02})_{0.95}Pb(I_{0.98}Br_{0.02})_3$ (where MA and FA denote methylammonium and formamidinium, respectively) and $FcTc_2$ surface treatment. Time-of-flight secondary ion mass spectrometry (TOF-SIMS) in Fig. 1B demonstrates that most of the $FcTc_2$ was located on the surface of the perovskite film. We used x-ray diffraction (XRD), top-view SEM, and UV-vis absorption spectroscopy measurements to study the crystallinity, morphology, and optical absorption of perovskite films with and without $FcTc_2$ treatment (figs. S4 to S6). All of the samples showed

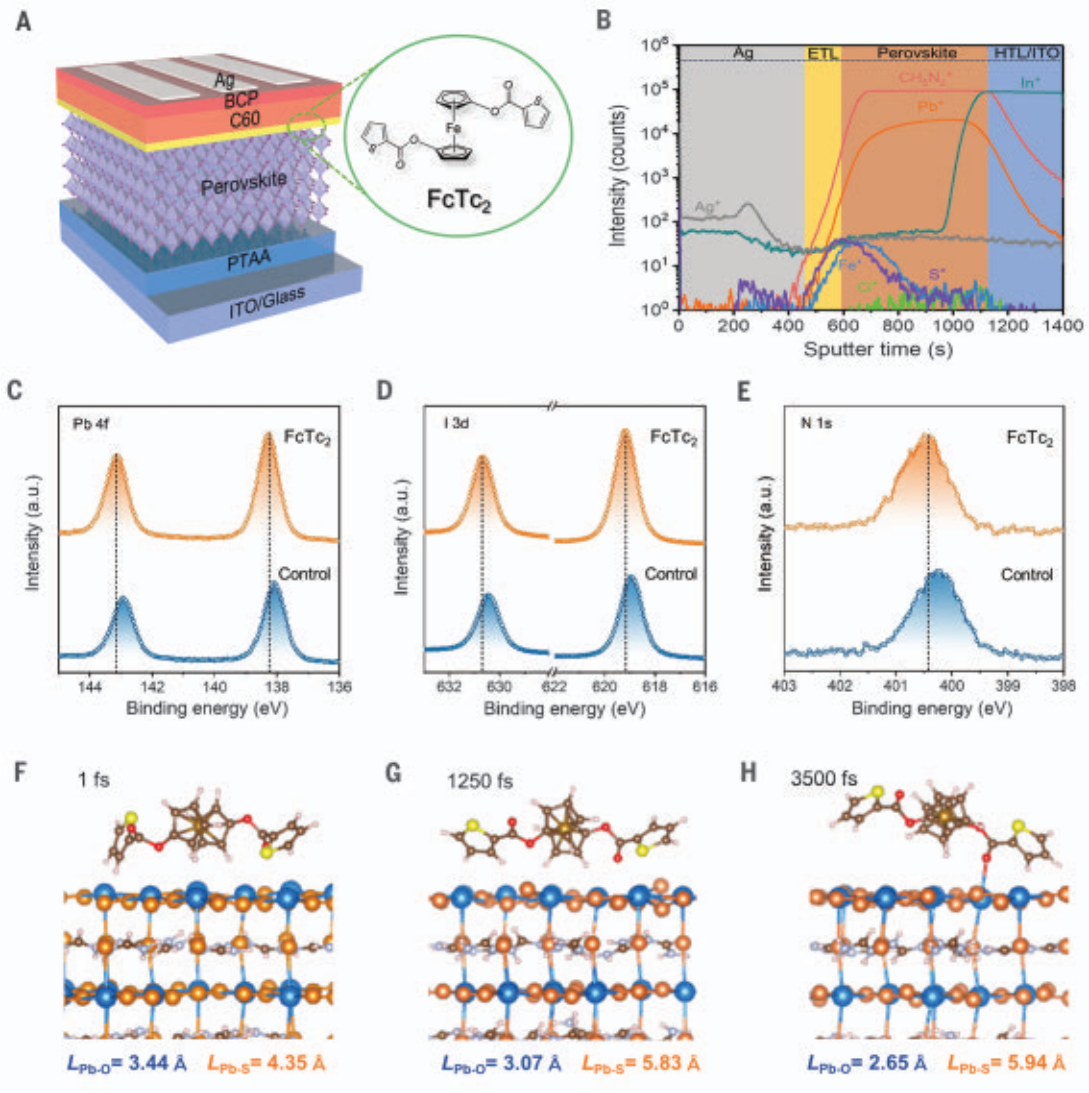


Fig. 1. Metal halide perovskite surface functionalized by $FcTc_2$. (**A**) Schematic illustration of inverted PSC based on $FcTc_2$ as the interface functionalization material. BCP, bathocuproine; ITO, indium tin oxide. (**B**) TOF-SIMS characterization of the PSC. (**C** to **E**) XPS characterization of Pb element (C), I element (D), and N element (E) on the perovskite film with and without $FcTc_2$. a.u., arbitrary units. (**F** to **H**) Molecular dynamics simulations of the interaction between perovskite and $FcTc_2$. *L*, bond length.

no obvious changes, which indicates that $FcTc_2$ does not affect the crystallization and light-harvesting properties of perovskite films.

The interaction of $FcTc_2$ with the perovskite was studied by comparing the x-ray photoelectron spectroscopy (XPS) measurements on the pristine and $FcTc_2$-treated perovskite films (Fig. 1, C to E). The binding energies corresponding to the Pb 4f, I 3d, and N 1s core levels of the $FcTc_2$-treated perovskite films all shifted marginally to higher values compared with the control sample, suggesting enhanced binding of both anions and cations on the perovskite surface, which could be caused by strong binding between surface ions and $FcTc_2$ (*17*).

Density functional theory (DFT) simulation analyses were performed to study the interaction between the perovskite surface and $FcTc_2$ molecules. We chose the (001) PbI_2-terminated perovskite surface as a model because it has been proven to be stable and to have the lowest energy configuration. Starting from the ordered interface, we observed enhanced bonding of O from $FcTc_2$ with Pb from the perovskite surface within a few picoseconds (Fig. 1, F and G). With the interfacial rearrangement, the molecular dynamics reach a stable equilibrium state in which the bond length of Pb-O was simulated to be 2.65 Å (Fig. 1H). The electrostatic potential (ESP) analysis (fig. S7) indicated a high electronegativity (−29.79 kcal mol^{-1}) of O in $FcTc_2$, driving the formation of stable Pb-O bonds and substantially enhancing the electrostatic attraction between the perovskite and $FcTc_2$ interface. XPS analysis combined with DFT simulation proves that there is a strong interaction between perovskite and $FcTc_2$, which is beneficial for both passivation of surface defects and stabilization of surface components in perovskite (*7*, *18*).

To study the effect of $FcTc_2$ on the electrical properties of perovskite films, we conducted kelvin probe force microscopy (KPFM) measurements to examine the surface potential of the films (Fig. 2, A and B). The perovskite film functionalized by $FcTc_2$ exhibited a decreased contact potential (~50 mV) relative to that of the control sample, which suggests direct interaction and surface charge transfer between $FcTc_2$ and perovskite (*18*). Moreover, $FcTc_2$-functionalized perovskite displayed a smaller potential distribution and surface potential difference (~150 mV) compared with that of the control sample (~250 mV). A uniform distribution of surface contact potential was beneficial for effective charge carrier extraction to prevent nonradiative recombination (*19*, *20*). The carrier kinetics characterization (figs. S8 and S9) on the perovskite-ETL interface further verifies the accelerated electron extraction with $FcTc_2$ modification, which can be attributed to the ferrocene and thiophene units (*21*, *22*).

Time-resolved photoluminescence (TRPL) spectra were measured to evaluate the nonradiative recombination of perovskite films (Fig. 2C). The carrier lifetime nearly doubled from 1166.74 to 2159.22 ns after incorporating $FcTc_2$ (the fitted carrier lifetime is summarized

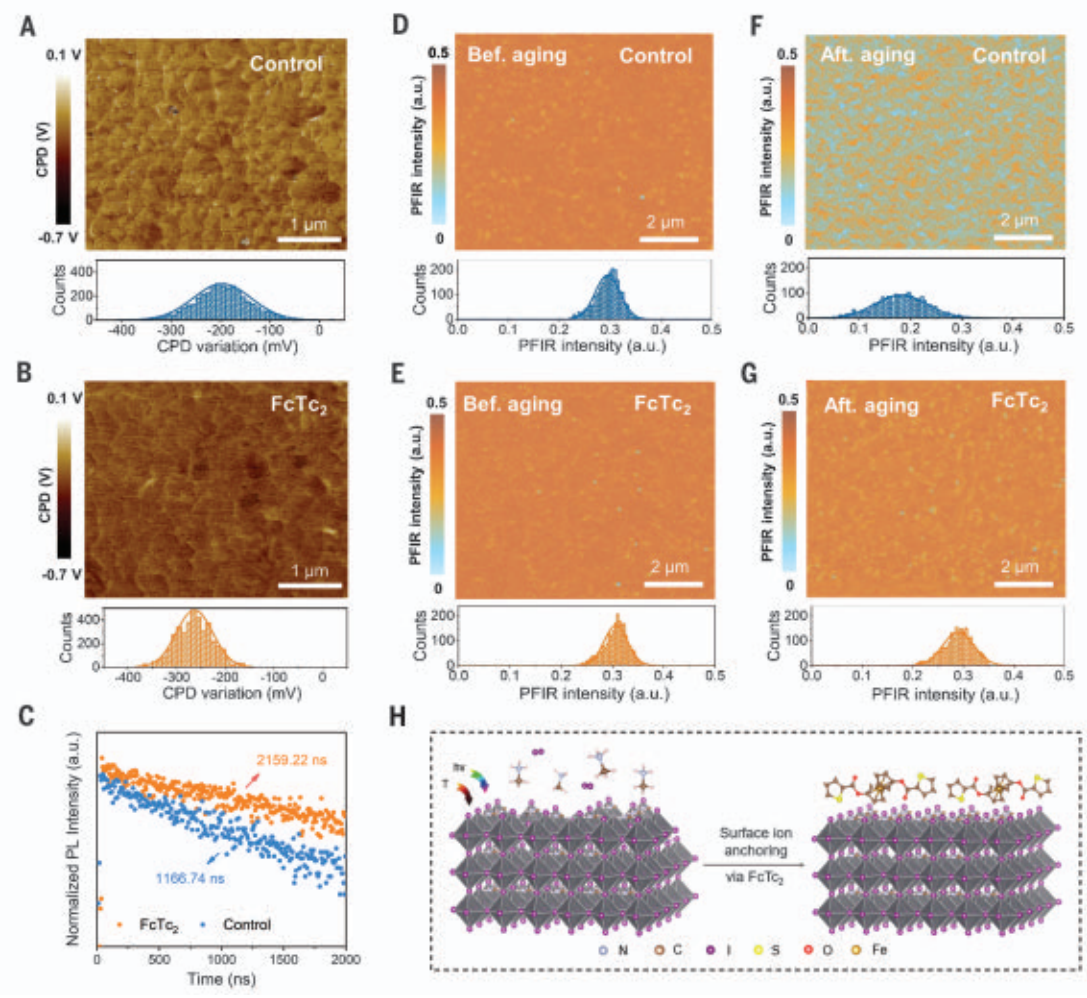


Fig. 2. Characterization of film surface properties. (**A** and **B**) Surface potential images were obtained by scanning KPFM of the control and $FcTc_2$-treated perovskite films. The statistical potential distributions of film surfaces are shown at the bottom. CPD, contact potential difference. (**C**) TRPL characterization of control and $FcTc_2$-treated perovskite films. PL, photoluminescence. (**D** to **G**) PFIR microscopy at an infrared (IR) frequency of 1480 cm^{-1} (which is resonant with the C-N stretching absorption of MA^+ ions) of control [(D) and (F)] and $FcTc_2$-modified perovskite films [(E) and (G)] before (Bef.) and after (Aft.) illumination at 85°C for 1000 hours. (**H**) Schematic illustration of the stabilization of surface ions by $FcTc_2$. T, temperature; hv, photon energy.

in table S1), which is consistent with enhanced steady-state photoluminescence intensity (fig. S10), indicating reduced numbers of nonradiative recombination centers from the surface defects. Additionally, the space charge–limited current (SCLC) measurements (fig. S12) further confirm the decreased defect density through $FcTc_2$ modification.

In triple-cation mixed-halide perovskite, the chemically reactive components, such as MA^+ and I^- at the perovskite surface, can easily volatilize and migrate through photo- and thermal effects. The surface trap states generate decreased photovoltaic performance degradation (*23*, *24*). To estimate the effect of $FcTc_2$ on perovskite stability, the MA^+ cations of the control and $FcTc_2$-functionalized perovskite films were probed by peak force infrared (PFIR) microscopy under illumination and heat conditions (*20*). The Fourier transform infrared (FTIR) spectroscopy spectra of the perovskite films confirmed that the signal of the MA ions in perovskite is distinct and readily resolved (fig. S13). PFIR mapping showed that the intensity and distribution of MA^+ cations in the $FcTc_2$-treated sample were well maintained after aging for 1000 hours (Fig. 2, D and F), whereas the control sample exhibited substantial reduction of intensity and broadening of distribution of the MA signal (Fig. 2, E and G). These results indicate that $FcTc_2$ can prevent surface ions from migration to produce a more uniform and stable surface component distribution (Fig. 2H) (*25*). By contrast, ion migration and volatilization were more prone to occur in the control films, resulting in increased surface defects and affecting the operational stability of perovskite devices (fig. S14). Measuring the perovskite films after >1000 hours separately under continuous 1-sun illumination or heating at 85°C also confirmed the stabilization effects of $FcTc_2$ on perovskite (figs. S15 and S16).

Figure 3A shows the current density–voltage (*J-V*) curves for control and $FcTc_2$-functionalized devices under AM1.5G simulated solar illumination, in which the concentration of $FcTc_2$ was optimized to be 1.0 mg ml^{-1} to obtain the best performance (figs. S17 and S18 and table S2). The control device exhibited a maximum PCE of 23.0%, with an open-circuit voltage (V_{OC}) of 1.13 V, a short-circuit current density (J_{SC}) of 25.25 mA cm^{-2}, and a fill factor (FF) of 80.45%. Compared with the control device, the $FcTc_2$-functionalized device exhibited an enhanced PCE of 25.0%, with an increased V_{OC} of 1.184 V, a J_{SC} of 25.68 mA cm^{-2}, and an FF of 82.32% with a low hysteresis (fig. S19). Corresponding external quantum efficiency (EQE) spectra (Fig. 3B) yielded integrated J_{SC} values

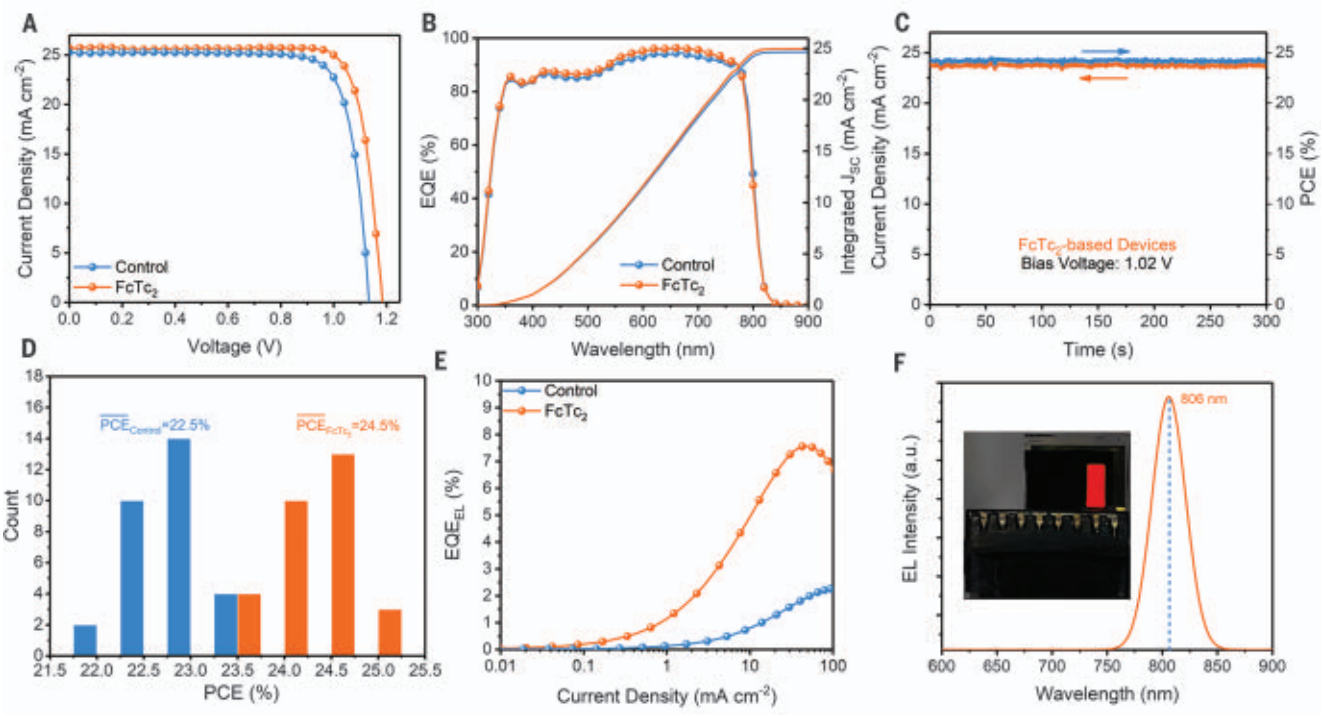


Fig. 3. Photovoltaic performance of PSCs. (**A**) *J-V* curves of the best-performing devices with and without $FcTc_2$. (**B**) EQE spectra and integrated current densities of the best-performing devices with and without $FcTc_2$. (**C**) Stabilized power output at the MPP for the best-performing PSCs treated with $FcTc_2$. (**D**) Histogram of the PCE values among 30 devices with and without $FcTc_2$. (**E**) EQE_{EL} values of the PSCs with and without $FcTc_2$ operating in light-emitting diode (LED) mode under different voltages. (**F**) EL spectra of the $FcTc_2$-based PSC operating as an LED.

with a small variation from the values obtained from *J-V* measurements. $FcTc_2$-treated devices were also measured at the maximum power point (MPP) to obtain a stabilized photocurrent of 23.68 mA cm^{-2} and a stabilized PCE of 24.2% (Fig. 3C). One of the best-performing devices was validated by an independent solar cell–accredited laboratory (National Institute of Metrology, China) for certification, where a PCE of 24.3% (with V_{OC} = 1.179 V, J_{SC} = 25.59 mA cm^{-2}, and FF = 80.60%) was confirmed (fig. S20), which is the highest certified efficiency among all inverted PSCs to date (table S3). The PCEs also exhibited a satisfying reproducibility, with an average PCE of 22.5% for the control device and 24.5% for the $FcTc_2$-based device (Fig. 3D). The particular effectiveness of $FcTc_2$ is further verified in figs. S21 to S24 and table S4.

Additionally, we conducted quantitative analysis of the photovoltage loss (V_{OC} loss) for the control and $FcTc_2$-treated devices according to detailed balance theory (*26*). The EQE_{EL} values of 1.5 and 7.0% for the control device and the device with $FcTc_2$, respectively, could be obtained from electroluminescence (EL) spectra (Fig. 3, E and F), leading to 108.57 and 68.75 mV of ΔV_3 (V_{OC} loss from the nonradiative recombination), respectively. The relative lower ΔV_3 suggests that $FcTc_2$ suppresses nonradiative recombination as an interfacial modifier. The values of the three components of V_{OC} loss (ΔV_1, ΔV_2, and ΔV_3)

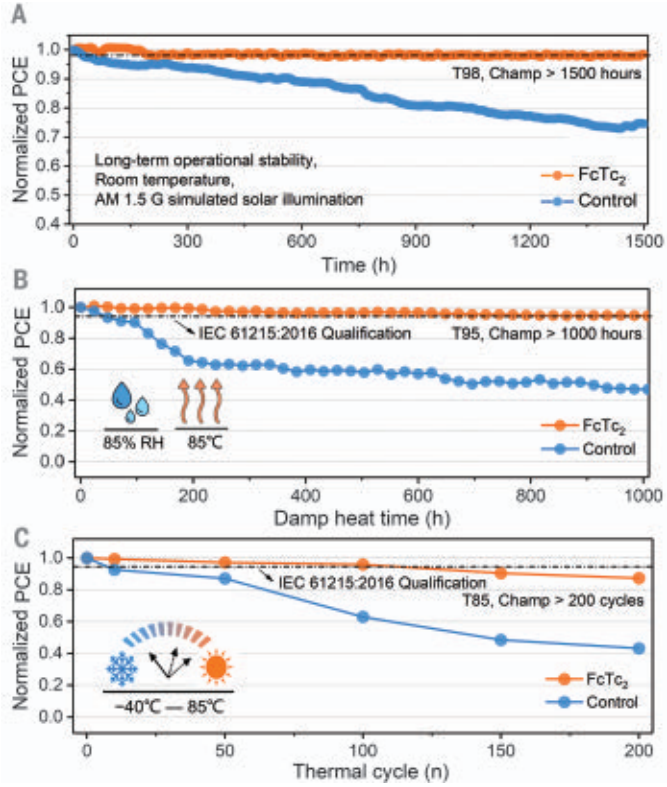


Fig. 4. Stability measurements of PSCs. (**A**) Normalized PCE of the unencapsulated PSCs with or without $FcTc_2$ measured at MPP under continuous 1-sun illumination in an N_2 atmosphere and at room temperature. (**B** and **C**) Normalized PCE of the PSCs with or without $FcTc_2$ measured following the IEC61215:2016 standards. (B) Encapsulated devices stored in 85% RH and 85°C in the dark. (C) Encapsulated devices stored in −40°C (15-min dwell) to 85°C (15-min dwell), ramp rate of 100°C $hour^{-1}$.

are calculated in the supplementary text and summarized in table S5. A V_{OC} loss of 363 mV is one of the lowest values reported among inverted PSCs (*16*, *26*).

To investigate the effect of $FcTc_2$ functionalization on device stability, the efficiency evolution under various conditions was monitored. We first examined the long-term operational stability of unencapsulated devices by loading at MPP voltage under continuous 1-sun illumination in an N_2 atmosphere (Fig. 4A). The $FcTc_2$-functionalized device retained its initial PCE in the first 200 hours and only exhibited a decay of <2% after >1500 hours. In comparison, the control device decreased to 72% of its initial PCE. We further measured the stability of unencapsulated devices under heat and ambient conditions. As shown in figs. S27 and S28, the performance of the control devices dropped to <80% of the initial efficiency after >800 hours. By contrast, the $FcTc_2$-functionalized devices showed a T98 (time to 98% of initial PCE) of 2000 hours under an ambient environment and 1500 hours under continuous heating. Because the chemically reactive components (such as MA^+ and I^-) at the perovskite surface can readily volatilize and migrate through photo-, humidity, and thermal degradation (*27*, *28*), we inferred that $FcTc_2$ enhances stability through the formation of additional bonding with perovskite surface ions and prevents any easily mobile ions from migration.

Additionally, we conducted strict stability measurements following the IEC61215:2016 standard, which is the most-used international standard for mature photovoltaic technologies. As shown in Fig. 4B, the $FcTc_2$-modified devices exhibited a T95 of >1000 hours under the damp heat test (85°C and 85% RH) and thus successfully passed the main point of IEC61215:2016 qualification for damp and heat conditions. Moreover, under the cycle shocks of cold (−40°C) and heat (85°C) shown in Fig. 4C, >85% PCE was retained after 200 cycles for the $FcTc_2$-modified devices, outperforming the control devices (40% PCE retained after 200 cycles). Taken together, these data indicate that $FcTc_2$-functionalized PSC devices exhibit excellent efficiency and stability and have the potential to move toward commercialization and rival traditional silicon solar cells.

REFERENCES AND NOTES

1. National Renewable Energy Laboratory, "Best research-cell efficiency chart" (2022); www.nrel.gov/pv/cell-efficiency.html.
2. L. Meng, J. You, Y. Yang, *Nat. Commun.* **9**, 5265 (2018).
3. G. Yang *et al.*, *Nat. Photonics* **15**, 681–689 (2021).
4. Y. Rong *et al.*, *Science* **361**, eaat8235 (2018).
5. J.-P. Correa-Baena *et al.*, *Science* **358**, 739–744 (2017).
6. N. Li, X. Niu, Q. Chen, H. Zhou, *Chem. Soc. Rev.* **49**, 8235–8286 (2020).
7. S. Yang *et al.*, *Science* **365**, 473–478 (2019).
8. X. Zheng *et al.*, *Nat. Energy* **5**, 131–140 (2020).
9. S. Wu *et al.*, *Nat. Nanotechnol.* **15**, 934–940 (2020).
10. S. Wu *et al.*, *Joule* **4**, 1248–1262 (2020).
11. S. Chen *et al.*, *Science* **373**, 902–907 (2021).
12. S. Bai *et al.*, *Nature* **571**, 245–250 (2019).
13. A. Mei *et al.*, *Joule* **4**, 2646–2660 (2020).
14. D. Luo *et al.*, *Science* **360**, 1442–1446 (2018).
15. J. J. Yoo *et al.*, *Nature* **590**, 587–593 (2021).
16. F. Li *et al.*, *J. Am. Chem. Soc.* **142**, 20134–20142 (2020).
17. Y. Guo *et al.*, *Nat. Commun.* **12**, 644 (2021).
18. R. Wang *et al.*, *Science* **366**, 1509–1513 (2019).
19. M. I. Saidaminov *et al.*, *Nat. Mater.* **19**, 412–418 (2020).
20. N. Li *et al.*, *Science* **373**, 561–567 (2021).
21. J. T. DuBose, P. V. Kamat, *J. Phys. Chem. Lett.* **10**, 6074–6080 (2019).
22. J. Zhang *et al.*, *Nanoscale* **10**, 5617–5625 (2018).
23. A. N. Singh *et al.*, *Adv. Energy Mater.* **10**, 2000768 (2020).
24. L. K. Ono, S. F. Liu, Y. Qi, *Angew. Chem. Int. Ed.* **59**, 6676–6698 (2020).
25. D. S. Jakob *et al.*, *Angew. Chem. Int. Ed.* **59**, 16083–16090 (2020).
26. J. Wang *et al.*, *Nat. Commun.* **11**, 177 (2020).
27. Q. Wang *et al.*, *Energy Environ. Sci.* **10**, 516–522 (2017).
28. Z. Fan *et al.*, *Joule* **1**, 548–562 (2017).

ACKNOWLEDGMENTS

Z.Z. thanks A. K.-Y. Jen and C. S. Lee for help and support. **Funding:** This work was supported by the New Faculty Start-up Grant of the City University of Hong Kong (9610421); the Innovation and Technology Fund (ITS/095/20); the ECS grant (21301319) and a GRF grant (11306521) from the Research Grants Council of Hong Kong; the Natural Science Foundation of Guangdong Province (2019A1515010761); and the Imperial College London Frankland Chair Endowment. N.J.L. is grateful for a Royal Society Wolfson Research Merit award. **Author contributions:** Z.L., B.L., and X.W. contributed equally to this work. Z.Z. conceived the ideas and designed the project with N.J.L. Z.Z. and N.J.L. directed and supervised the research. Z.L. fabricated the devices, conducted the characterization, and analyzed the data. B.L. and X.W. also contributed to the characterization and data analyses. N.J.L. and S.A.S. designed and synthesized the materials. S.Z. conducted the DFT calculations. D.G. conducted the stability tests. Z.L., B.L., X.W., N.J.L., and Z.Z. drafted and finalized the manuscript. All the authors revised the manuscript. **Competing interests:** A patent application based on this work has been submitted by Imperial College London and City University of Hong Kong, led by N.J.L. and Z.Z. The authors declare no other competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper or the supplementary materials.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abm8566
Materials and Methods
Supplementary Text
Figs. S1 to S28
Tables S1 to S5
References (*29*–*44*)

21 October 2021; accepted 17 March 2022
10.1126/science.abm8566

FISHERIES

Rapid evolution in salmon life history induced by direct and indirect effects of fishing

Y. Czorlich[1,2,3,4], T. Aykanat[3], J. Erkinaro[2], P. Orell[2], C. R. Primmer[3,5]*

Understanding the drivers of evolution is a fundamental aim in biology. However, identifying the evolutionary impacts of human activities is challenging because of a lack of temporal data and limited knowledge of the genetic basis of most traits. Here, we identify the drivers of evolution toward maturity at an earlier age in Atlantic salmon through two types of fisheries-induced evolution acting in opposing directions: an indirect effect linked with harvest of a salmon prey species (capelin) at sea (selection against late maturation) and a direct effect due to net fishing in rivers (selection against early maturation). Because capelin are harvested as an aquaculture feed protein source, we hereby determine an indirect path by which salmon aquaculture may influence wild salmon populations.

Indirect ecological effects (e.g., a third species mediating the impact of one species on another) (*1*) and their evolutionary impacts have often been neglected in studies of natural populations. For instance, fishing can induce evolution of traits such as size and age at maturity in the target species (*2*, *3*) but may also induce larger, ecosystem-level changes, some of which may be indirect. However, cases demonstrating direct effects at the genetic level are rare, as are empirical examples of the indirect evolutionary impacts of fishing (*2*, *4*). Knowledge of indirect ecological and evolutionary effects is critical for properly evaluating the consequences of different fisheries management strategies (*5*).

Atlantic salmon (*Salmo salar*) have a complex life history, utilizing both freshwater and marine habitats, and thus affect and are affected by multiple ecosystems (*6*). Sea age is defined as the number of years an individual spends in a marine environment before returning to fresh water to spawn. In Atlantic salmon, sea age at maturity is an important life history trait that exhibits an evolutionary trade-off between survival and reproduction. Individuals that mature later are larger and have higher reproductive success (*7*) but run a higher risk of mortality before spawning. Age at maturity has been associated with a major effect locus in the genome region that includes the *vgll3* gene and

[1]University of Turku, Department of Biology, FI-20014 Turku, Finland. [2]Natural Resources Institute Finland (Luke), POB 413, FI-90014 Oulu, Finland. [3]University of Helsinki, Organismal & Evolutionary Biology Research Programme, POB 56, FI-00014 Helsinki, Finland. [4]Norwegian Institute for Nature Research (NINA), NO-7485 Trondheim, Norway. [5]University of Helsinki, Institute of Biotechnology, Helsinki Institute of Life Science (HiLIFE), POB 56, FI-00014 Helsinki, Finland.
***Corresponding author. Email: craig.primmer@helsinki.fi**

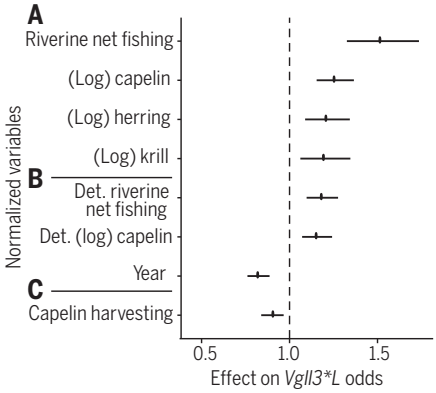


Fig. 1. Standardized regression coefficients for variables significantly associated with *vgll3*L* odds (later and larger maturation). The estimates come from the (**A**) initial quasi-binomial model, (**B**) detrended (Det) model, and (**C**) Monte Carlo Method for assessing mediation (i.e., indirect effects). The dotted line indicates no effect on *vgll3*L* odds. The error bars correspond to 95% CIs.

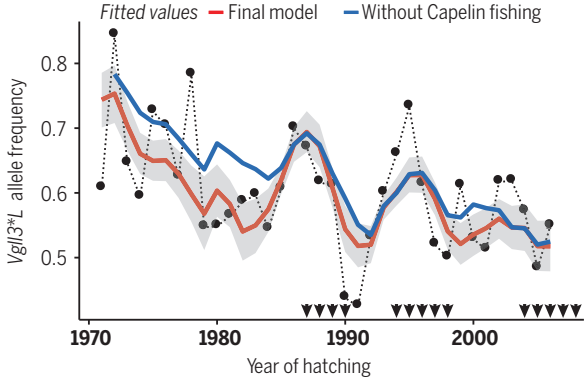


Fig. 2. Temporal changes in *vgll3*L* (later maturation) allele frequency. The black dotted line represents the observed data, the red line represents the allele frequency estimated from the final model including the direct and indirect effects of fishing, and the blue line represents the expected *vgll3*L* allele frequency assuming no capelin fishing (annual effect on returning salmon). Black arrows indicate years when there was a moratorium on capelin harvesting (harvest rate <0.5 %). As *vgll3* allele frequencies are based on hatch date, capelin fishing closures are expected to influence *vgll3* allele frequencies of salmon hatching a few years earlier than the fishing closures, as a result of salmon spending 3 to 5 years in fresh water before migrating to the sea. Shaded areas indicate 95% bootstrap intervals based on 3000 replicates (see materials and methods for details).

explains 40% of the variation in the trait (*8*). Adaptive evolution toward younger age at maturity at the *vgll3* locus over the course of 40 years has previously been demonstrated in a large Atlantic salmon population from the Teno River (Deatnu in Sámi, Tana in Norwegian) in northern Europe (*9*), but the environmental drivers of evolution in age at maturity were not identified.

We investigated ecological and environmental variables that potentially affect the relative fitness of salmon with different maturation ages (and therefore sizes) from the same river; these variables include fishing effort at sea, fishing effort in the river, sea temperature, and the average biomass of three key prey species [capelin (*Mallotus villosus*), herring (*Clupea harengus*), and krill; fig. S1] available during the years each salmon spent in the Barents Sea ecosystem. We used individual-based quasi-binomial generalized linear models (GLMs) to identify environmental variables linked with temporal variation in *vgll3* allele frequencies (*10*) and thereby age at maturity in a 40-year time series (1975–2014) consisting of 1319 individuals genetically assigned to the Tenojoki population (hereafter, Teno) in the middle reaches of the mainstem of the Teno river system (*9*). These analyses indicate that annual number of riverine net-fishing licenses (a proxy for annual fishing pressure) had the strongest effect of the abovementioned variables on *vgll3* allele frequency (Fig. 1). Notably, annual riverine fishing pressure was positively associated with the *vgll3*L* allele frequency (standardized regression coefficient [$\beta^* = 0.42$, $F_1 = 27.79$, and P value < 0.001], indicating that higher net fishing pressure in the river was associated with higher frequencies of the allele associated with later maturation in salmon and therefore larger sizes (table S1). We verified the results by detrending the data (using a residual regression approach) to account for potentially confounding factors creating temporal trends between the dependent and independent variables and thus avoid possible spurious associations (*11*). The association between annual net fishing license numbers and *vgll3* remained in the detrended model and included a significant and negative year effect, indicating that fishing pressure is also linked to annual *vgll3* allele frequency changes around the trend [$\beta^* = 0.17$, F(1) = 14.15, and P value < 0.001; Fig. 1 and table S1]. Capelin biomass in the Barents Sea was also positively associated with the frequency of the *vgll3* allele associated with later maturation and larger size (*vgll3*L*) in salmon in both the normal [$\beta^* = 0.23$, F(1) = 20.77, and P value < 0.001] and detrended models [$\beta^* = 0.14$, F(1) = 10.64, and P value = 0.001; table S1]. Herring and krill biomass also had a significant effect on *vgll3* allele frequencies in a similar direction to capelin (Fig. 1 and table S1); however, these effects did not remain significant in the detrended model. There was little evidence for associations between the other variables and *vgll3* in the GLM (rod fishing licenses in the river, Barents Sea temperature, and net fishing at sea; table S1).

Analyses at the phenotypic level indirectly support the influence of prey biomass and fishing on Teno salmon fitness. After controlling for the *vgll3* sex-specific genetic effects, a multinomial model indicates that the proportion of older, later-maturing Atlantic salmon in the river increases as capelin, herring, and krill biomass increases (table S2). This result is expected if the abundance of these prey species is positively associated with salmon survival at sea, as only those that return to the river are sampled. However, it can also reflect plastic changes in maturation probabilities. This model also shows that a higher number of riverine net-fishing licenses is associated with a higher proportion of late-maturing salmon observed at the end of the fishing season (table S2), which is expected if net fishing targeted preferentially small, early-maturing salmon. Given that sea age at maturity of Teno females is considerably higher on average than that of males and therefore time spent in the marine environment is longer (2.8 years for females versus 1.5 years for males), environmental conditions that strongly affect marine survival are also expected to influence the sex ratio of adults returning from their migration to spawn. In accordance with this prediction, a binomial model showed that the proportion of returning females increased with prey biomass and riverine net fishing (fig. S2 and table S3). The effects of riverine net fishing on sex ratio and age at maturity probability, however, were not significant in the detrended regressions (tables S2 and S3).

Forage fishes such as capelin play important roles in marine ecosystems by enabling energy transfer between upper trophic levels (predators such as large fish, seabirds, and mammals) and lower trophic levels (plankton) (*12*). In the Barents Sea, the capelin stock experienced several substantial collapses over the course of the 40-year study period (fig. S3) as a result of overexploitation in commercial fisheries combined with predation by herring and cod (fig. S4) (*13*). We therefore quantified the potential indirect effects of capelin harvesting (while accounting for other ecosystem interactions) on *vgll3* allele frequency dynamics using a multispecies Gompertz model developed by Langangen *et al.* (*14*) (figs. S1, S3, and S4). The analysis indicated a significant indirect effect of capelin harvest rate on evolution in age at maturity in Teno salmon, with a 30% decrease in the *vgll3*L* allele odds per harvest rate unit [Monte Carlo

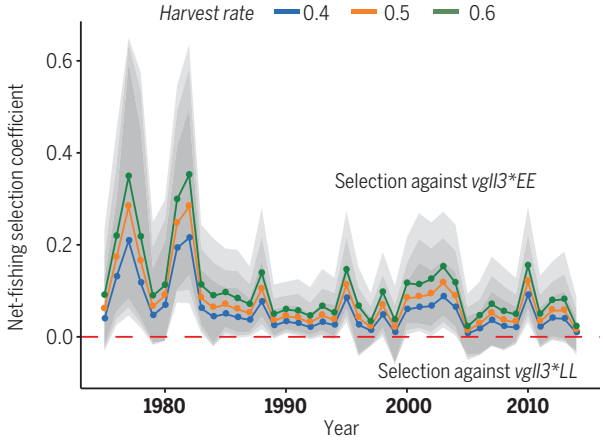


Fig. 3. Predicted variation of net fishing–induced selection over 40 years in Teno salmon as a function of harvest rate. The selection coefficient corresponds to the difference in relative survival between *vgll3*EE* and *vgll3*LL*. Calculations are based on capture probabilities (with harvest rates ranging from 40 to 60% of returning individuals) as a function of length, using sonar data in 2018 and 2019. It accounts for the sex-specific *vgll3* association with length during the study period, temporal variation in sex ratio, and relative use of weirs, drift nets, and gill nets. The shaded area represents 95% CIs based on 6000 iterations.

method for assessing mediation; confidence interval (CI): 95% = (0.116 to 0.471)]. The evolutionary effect of capelin harvest was strongest during the early years of the time series, likely because of repeated capelin fishery closures and fishing effort reduction after the capelin stock collapses (Fig. 2 and fig. S5A). This is the same time period when most of the evolutionary changes in age at maturity for males occurred (*9*). However, by reducing the *vgll3*L* allele frequency in adults that return to reproduce, the indirect fishing effect (Fig. 2) is also expected to influence the allele frequency of offspring in the next generation due to the lower *vgll3*L* allele frequency in the breeding adults.

The direction of selection pressure in riverine net fishing may seem counterintuitive, as net fishing is generally thought to select against large individuals (*15, 16*). However, net fishing in the Teno river includes the use of multiple net types with different selectivity: weirs, gill nets, and drift nets (*17*). To estimate the size selectivity of different fishing gear types, we compared the size distribution of salmon caught with each type during the fishing period in June and July of 2018 and 2019 with the size distribution of ascending salmon obtained from a sonar count in the river during the same period. This analysis indicates that weir fishing selects against the early maturation allele (*vgll3*E*) by capturing a higher proportion of smaller and earlier-maturing individuals, whereas drift-net and gill-net fishing select against the *vgll3*L* allele (figs. S6 and S7). Discussions with Indigenous Sámi fishers with decades of net-fishing experience revealed that the tendency for weirs to preferentially capture younger, smaller individuals could potentially arise from several factors, including the positioning of weirs primarily in shallower waters, smaller mesh size, and/or the timing of use during the fishing season (*18*). The positive effect of riverine net license number (including all net types) on *vgll3*L* allele frequency may thus be explained by the predominant use of weirs, the most selective fishing method (representing 54% of net catches on average). Over the time series, net-fishing effort declined considerably as a result of stricter fishing regulations in addition to a decrease in the number of Indigenous fishers using weirs (for example, both net-fishing licenses and number of weirs were >30% lower in 2014 versus 1976; fig. S5B). In addition to a decline in net-fishing effort, temporal changes in other factors may have resulted in a decrease in size selectivity of net fishing, including changes in the relative use of different fishing gear and changes in salmon size distribution and sex ratio (Fig. 3). Consequently, net-fishing selective pressure against the vgll3*E allele is expected to have declined during the time series, which has likely resulted in improved riverine survival of early-maturing salmon relative to older individuals. Overall, the extra mortality of late-maturing *vgll3*L* individuals at sea was not sufficiently compensated by their size-induced survival advantage as a result of riverine net fishing (decreasing over time) nor their reproductive advantage due to larger size (*7*), thus resulting in an overall fitness loss and leading to the observed overall decrease in *vgll3*L* allele frequency (Fig. 2).

Using an ecosystem-level approach, we provide an example in which both direct and indirect fishing effects have contributed to fisheries-induced evolution. Minimizing this type of evolution is generally recommended, as adaptation of populations to fishing can hinder adaptation to their natural environment and may be costly in the long run (*19–21*). However, ongoing socioeconomic issues may limit opportunities for major changes in fishing regulations to alter harvest rates to minimize fisheries-induced evolution (*21*). In the Teno river valley, for instance, salmon fishing is a key source of income for locals through fishing tourism as well as the foundation of the Indigenous Sámi culture and identity, and restrictive fisheries regulations can be controversial (*22*). The varying selectivity of different net types may provide a means to manage the selection pressures exerted by fishing on the different ages at maturity and genotypes—for example, by regulating the relative use of different types of fishing gear. Collection of additional data such as additional years of genetic stock assignment and sonar counting, as well as the development of models coupling population dynamics and evolutionary genetics, would be required before implementation. Such models would also help in identifying selection driven by factors such as fishing at a constant harvest rate that cannot be detected using regressions; such models would also help in determining the impact of fisheries-induced selection on population growth rate. Commercial harvest of capelin appears to have indirectly induced evolution of Atlantic salmon age at maturity toward younger, smaller individuals. Several studies have noted the potential for fisheries to induce ecological effects beyond the target species (*23*). However, evolutionary effects have not been demonstrated earlier (*2*, *4*). Our study also has implications of a more applied nature: ~90% of forage fish catch is used for fish oil and fishmeal to feed farmed animals (*12*) and salmonid fish aquaculture is the fourth-highest consumer (*24*). In 2012, for instance, 75,800 tons of capelin were used in Norwegian aquaculture salmon feed (representing 15% of the marine ingredients) (*25*). Our results therefore identify an indirect path by which Atlantic salmon aquaculture can affect wild populations of the same species and emphasize the importance of identifying alternative, sustainable protein sources for the aquaculture industry.

REFERENCES AND NOTES

1. M. R. Walsh, *Trends Ecol. Evol.* **28**, 23–29 (2013).
2. M. Heino, B. Díaz Pauli, U. Dieckmann, *Annu. Rev. Ecol. Evol. Syst.* **46**, 461–480 (2015).
3. A. Kuparinen, J. Merilä, *Trends Ecol. Evol.* **22**, 652–659 (2007).
4. J. A. Hutchings, A. Kuparinen, *Fish Fish.* **21**, 453–464 (2020).
5. L. B. Crowder *et al.*, *Annu. Rev. Ecol. Evol. Syst.* **39**, 259–278 (2008).
6. T. Forseth *et al.*, *ICES J. Mar. Sci.* **74**, 1496–1513 (2017).
7. K. B. Mobley *et al.*, *Sci. Adv.* **5**, eaav1112 (2019).
8. N. J. Barson *et al.*, *Nature* **528**, 405–408 (2015).
9. Y. Czorlich, T. Aykanat, J. Erkinaro, P. Orell, C. R. Primmer, *Nat. Ecol. Evol.* **2**, 1800–1807 (2018).
10. D. J. Schaid, J. P. Sinnwell, G. D. Jenkins, *Hum. Hered.* **74**, 71–82 (2012).
11. M. H. Graham, *Ecol. Soc. Am.* **84**, 2809–2815 (2003).
12. E. K. Pikitch *et al.*, *Fish Fish.* **15**, 43–64 (2012).
13. D. Ø. Hjermann, G. Ottersen, N. C. Stenseth, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 11679–11684 (2004).
14. Ø. Langangen *et al.*, *Glob. Change Biol.* **23**, 283–292 (2017).
15. W. E. Ricker, *Can. J. Fish. Aquat. Sci.* **38**, 1636–1656 (1981).
16. N. W. Kendall, T. P. Quinn, *Ecol. Appl.* **22**, 804–816 (2012).
17. J. Erkinaro *et al.*, *Can. J. Fish. Aquat. Sci.* **76**, 42–55 (2019).
18. A. Holmberg, N. A. Holmberg, Teno Valley Sámi, Traditional net fishing practices (2022).

19. M. Heino *et al*., *ICES J. Mar. Sci.* **70**, 707–721 (2013).
20. M. R. Walsh, S. B. Munch, S. Chiba, D. O. Conover, *Ecol. Lett.* **9**, 142–148 (2006).
21. J. J. Hard *et al*., *Evol. Appl.* **1**, 388–408 (2008).
22. J. Hiedanpää, J. Saijets, P. Jounela, M. Jokinen, S. Sarkki, *Environ. Manage.* **66**, 1039–1058 (2020).
23. R. L. Naylor *et al*., *Nature* **405**, 1017–1024 (2000).
24. H. E. Froehlich, N. S. Jacobsen, T. E. Essington, T. Clavelle, B. S. Halpern, *Nat. Sustain.* **1**, 298–303 (2018).
25. T. Ytrestøyl, T. S. Aas, T. Åsgård, *Aquaculture* **448**, 365–374 (2015).
26. Y. Czorlich, T. Aykanat, J. Erkinaro, P. Orell, C. R. Primmer, Data and custom codes from "Rapid evolution in salmon life-history induced by direct and indirect effects of fishing", version 1.00, Zenodo (2022); https://doi.org/10.5281/zenodo.5901803.

ACKNOWLEDGMENTS

We thank the fishers who collected scales and phenotypic information and the people who organized and read the scales, especially J. Kuusela and J. Haantie. We also thank J.-P. Pohjola for sonar data reading and M. Länsman for compiling the detailed river catch and effort data. A. Holmberg and N. A. Holmberg are thanked for sharing their vast knowledge of traditional Sámi fishing practices. We also thank E. Eriksen for providing corrected krill biomass data and Ø. Langangen for sharing codes and information on the multispecies Gompertz model. Four anonymous reviewers are thanked for constructive comments on earlier versions. **Funding:** This project received funding from the Academy of Finland (project 284941, 286334, 314254, 314255, 327255, and 325964) as well as from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement 742312). Part of Y.C.'s salary was funded by the Norwegian Research Council (projects No. 275862 EcoEvoGene and 280308 SeaSalar). Y.C. is also affiliated with the Center for Biodiversity Dynamics (CBD, NTNU). **Author contributions:** J.E. and P.O. coordinated the collection of scale samples and river fisheries data; C.R.P., Y.C., T.A., and J.E. designed the study; Y.C. performed the lab work and analyzed the data; Y.C., C.R.P., and T.A. wrote the manuscript and all authors contributed to its revision. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The datasets generated for the current study and the custom codes are available in a public data repository that becomes accessible upon publication (*26*). All other data needed to evaluate the conclusions in the paper are present in the paper or the supplementary materials.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abg5980
Materials and Methods
Figs. S1 to S8
Tables S1 to S4
References (*27*–*61*)

26 January 2021; accepted 14 February 2022
Published online 24 February 2022
10.1126/science.abg5980

CHANGE
YOUR JOB
AND YOU
JUST MIGHT
CHANGE
THE WORLD.
Find your next job at ScienceCareers.org
The relevance of science is at an all-time high these days. For anyone who's looking to get ahead in —or just plain get into— science, there's no better, more trusted resource or authority on the subject than *Science* Careers. Here you'll find opportunities and savvy advice across all disciplines and levels. There's no shortage of global problems today that science can't solve. Be part of the solution.
ScienceCareers
FROM THE JOURNAL SCIENCE
AAAS

Who's the top employer for 2021?

Science Careers' annual survey reveals the top companies in biotech & pharma voted on by *Science* readers.

Read the article and employer profiles at sciencecareers.org/topemployers

Science 2021 TOP EMPLOYER

tifr

TATA INSTITUTE OF FUNDAMENTAL RESEARCH
Homi Bhabha Road, Colaba, Mumbai 400 005
An Autonomous Institute of the Department of Atomic Energy, Government of India, and a Deemed University

Invites applications / nominations for the post of

CENTRE DIRECTOR,

NATIONAL CENTRE FOR BIOLOGICAL SCIENCES, BANGALORE

From eminent academics / researchers / scientists with high academic achievement and some experience of administration

For details please visit TIFR Website
http://www.tifr.res.in/positions

Last date for receipt of applications is **June 30, 2022**.

What's Your Next Career Move?

From networking to mentoring to evaluating your skills, find answers to your career questions on *Science* Careers

To view the complete collection, visit **ScienceCareers.org/booklets**

ScienceCareers
FROM THE JOURNAL SCIENCE AAAS

By **Kevin F. Boehnke**

Nurturing my ambitions

"You should apply!" my mentor exclaimed as they slid the job description across the table. It was for a tenure-track appointment—a position that had been my driving ambition—in the department where I had completed my Ph.D. Professionally, the timing was perfect: My postdoctoral training had expanded my skill set and I had recently received a great score on a grant application, poising me to return to my old digs with a unique niche and the proven ability to fund it. Personally, however, I wasn't sure. My two children were both under 3 years old and my current position—a research faculty appointment in a supportive team science environment—provided much-needed stability and flexibility. It was time to truly examine my ambitions.

For the previous year, I had been wrestling with a nebulous but inescapable sense of unease. My work was meaningful and engaging, but something felt off. The wheels in my mind were always turning, mapping out future studies, career advancement, and policy consequences of my research. But instead of feeling invigorated and stimulated, I would wake up at night sweating protocol details, leaving me impatient and distracted the next day. My infant experienced birds, flowers, and trees with unwavering gusto when we went on walks, but I was distracted, fixated on study design. I found myself neglecting my meditation and yoga practice—crucial routines for managing my chronic pain—to spend more time working. As time went on, so did my feeling of sleepwalking through life.

"I've learned to slow down, be gentle with myself, and prioritize relationships."

This discomfort crystallized one fine summer day when I heard a podcast about workism, the doctrine that employment should provide identity, purpose, and community. In a flash of embarrassed comprehension, I realized this term described much of how I had approached my post-Ph.D. life. I viewed studies, grants, collaborations, and impact as integral to my identity rather than to my job, equating personal worth with professional successes. I followed my self-care routines largely to ensure I was well enough to maximize academic productivity. Because there was always more work to be done, I felt I never measured up. Put simply, my professional mask was too sticky. I feared it might become permanently stuck.

I vowed to change course. I began small, working on taking deep breaths throughout the day to stay centered and paying more attention to my internal life. When I fell into an incessant thought loop about work, I would try to focus on a compassionate mantra: "I am enough. I am enough. I am enough." These practices felt grounding, but work kept derailing my efforts to return to yoga and meditation. I still wasn't quite where I wanted to be.

Then, the tenure-track opportunity arose, and I was torn. Although the changes I had made helped me understand the patterns I wanted to move away from, parts of me were still attracted to the position. I agonized for weeks, talking with anyone who would listen and tossing and turning at night.

I ultimately decided not to apply. It was tough to let the opportunity go, but I'm at peace with it. I haven't abandoned ambition; I have expanded it to encompass physical, emotional, spiritual, and familial goals. I recommitted to yoga and meditation through my personal practice and occasional teaching. My wife and I have devoted ourselves to putting down deeper roots and cultivating a loving, thoughtful community—a "village" for our family. I savor how I've learned to slow down, be gentle with myself, and prioritize relationships—all of which leaves me more fulfilled.

As my relationship with ambition evolves, I keep returning to the word "alignment," which my favorite definition describes as "proper positioning … of parts in relation to each other." I suspect that aligning my ambitions will be a lifelong effort. By working to channel my attention toward what really matters to me, I hope to find appropriate balance for the different life phases to come. ■

Kevin F. Boehnke is a research investigator at the University of Michigan, Ann Arbor. Send your career story to SciCareerEditor@aaas.org.

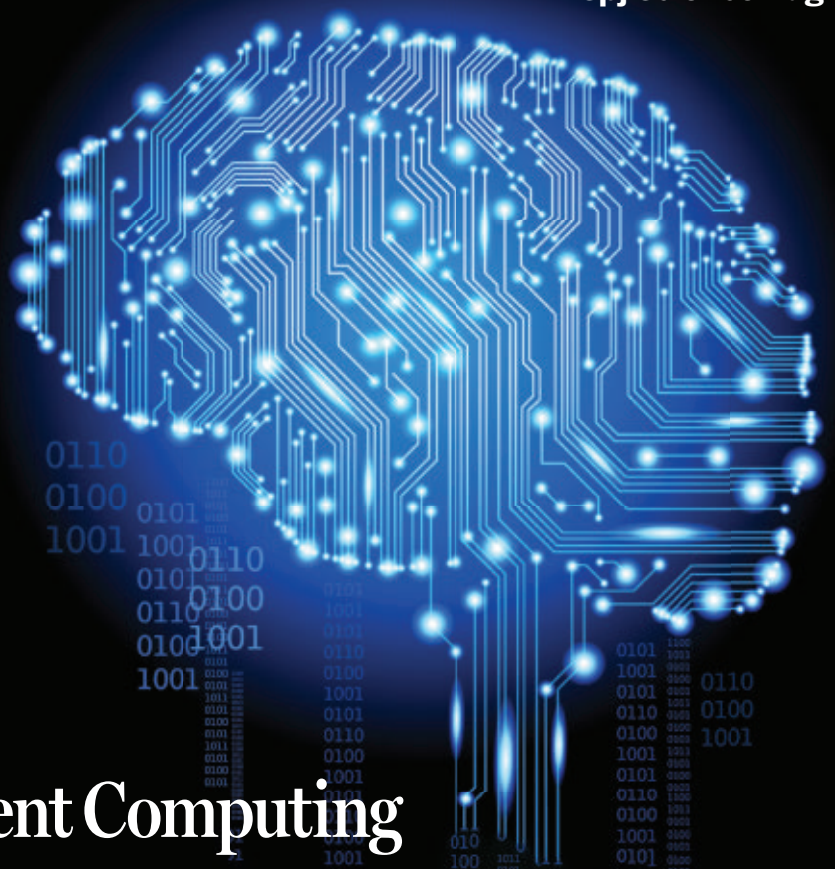
ILLUSTRATION: ROBERT NEUBECKER

CALL FOR PAPERS

SPJ
SCIENCE PARTNER JOURNALS

之江实验室
ZHEJIANG LAB

spj.sciencemag.org/icomputing



Intelligent Computing

Intelligent Computing is an Open Access journal published in affiliation with **Zhejiang Lab** and distributed by the **American Association for the Advancement of Science (AAAS)**. The journal aims to build a platform that promotes open communication and cooperation for global science and technology professionals in the field. *Intelligent Computing* presents the latest research outcomes and technological breakthroughs to boost the development of intelligent computing science and technology.

Submit your research to *Intelligent Computing* today!
Learn more at: **spj.sciencemag.org/icomputing**

The Science Partner Journals (SPJ) program was established by the American Association for the Advancement of Science (AAAS), the nonprofit publisher of the *Science* family of journals. The SPJ program features high-quality, online-only, open access publications produced in collaboration with international research institutions, foundations, funders and societies. Through these collaborations, AAAS expands its efforts to communicate science broadly and for the benefit of all people by providing top-tier international research organizations with the technology, visibility and publishing expertise that AAAS is uniquely positioned to offer as the world's largest general science membership society.

@SPJournals @SPJournals

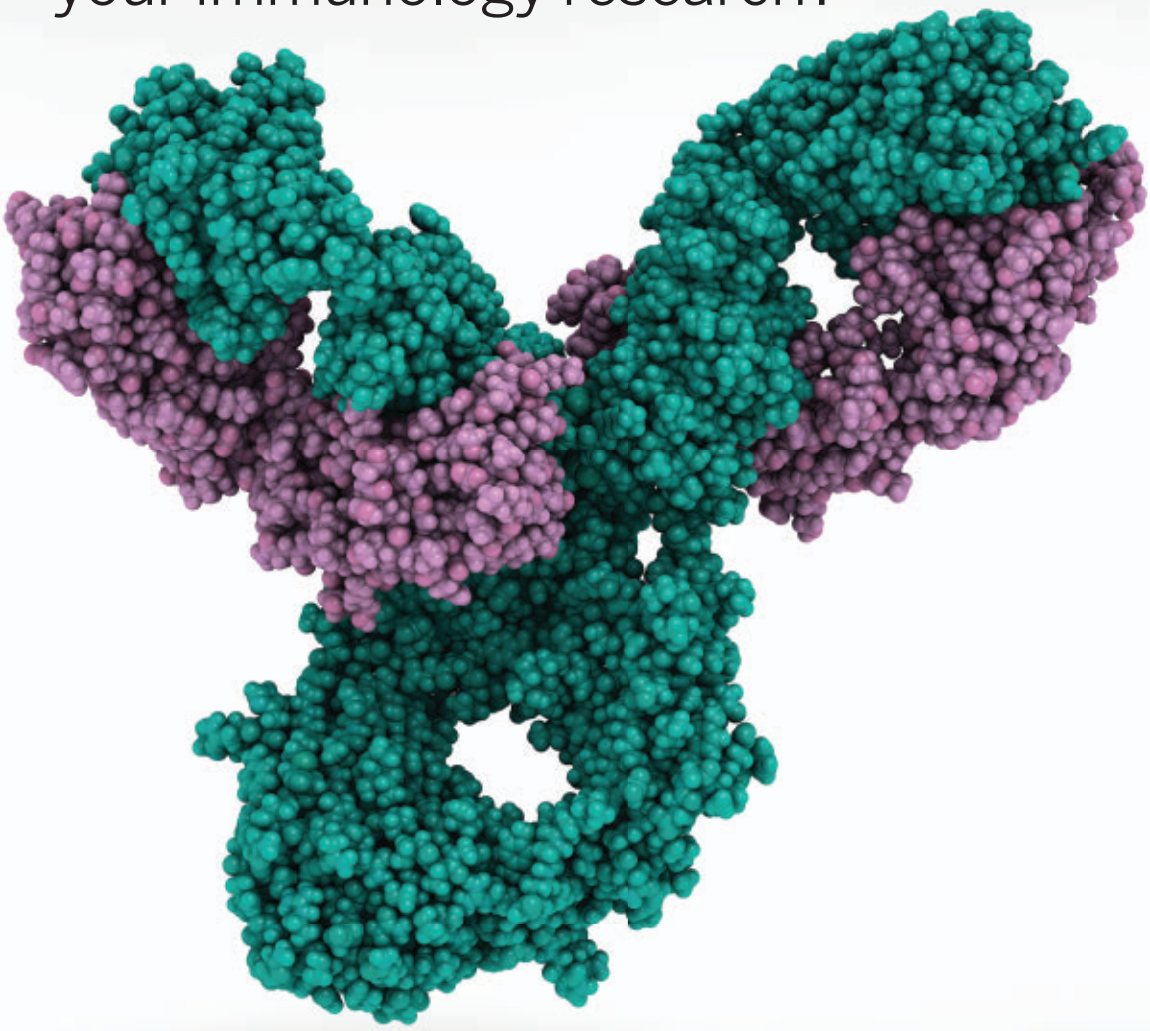
OPEN ACCESS

ARTICLE PROCESSING CHARGES WAIVED THROUGH 2024

Visit us at
booth **#239**

Antibodies Recognize Antigens...

How will you get **recognized** for your immunology research?



For more information, visit us at
science.org/journal/sciimmunol

Science Immunology | AAAS

Twitter: @SciImmunology Facebook: @ScienceImmunology