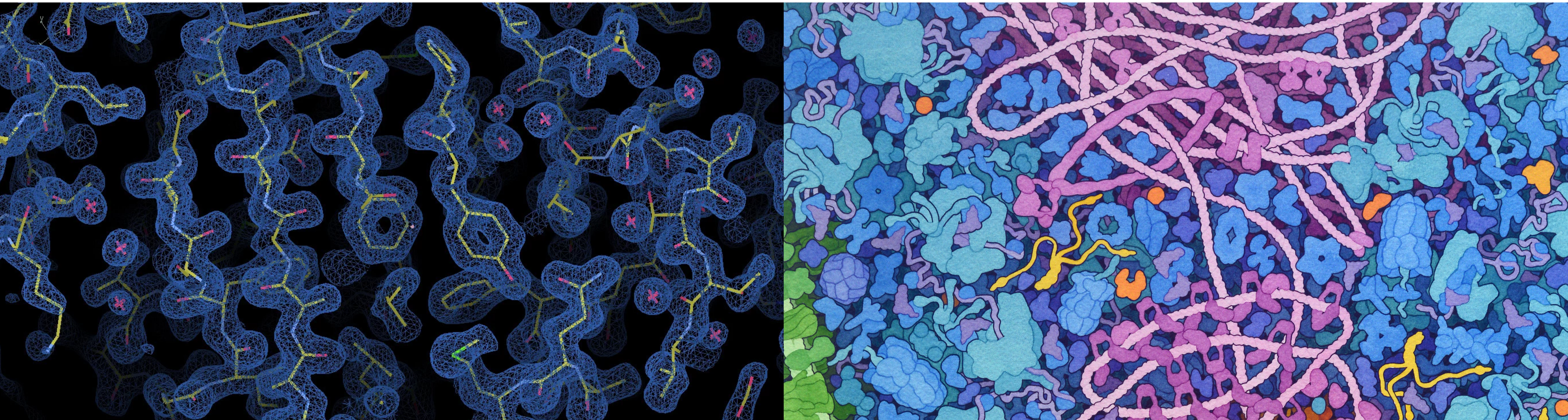


COS 597Y: Machine Learning for the Natural Sciences



Ellen Zhong
Department of Computer Science
Princeton University

Fall 2026
Course Introduction

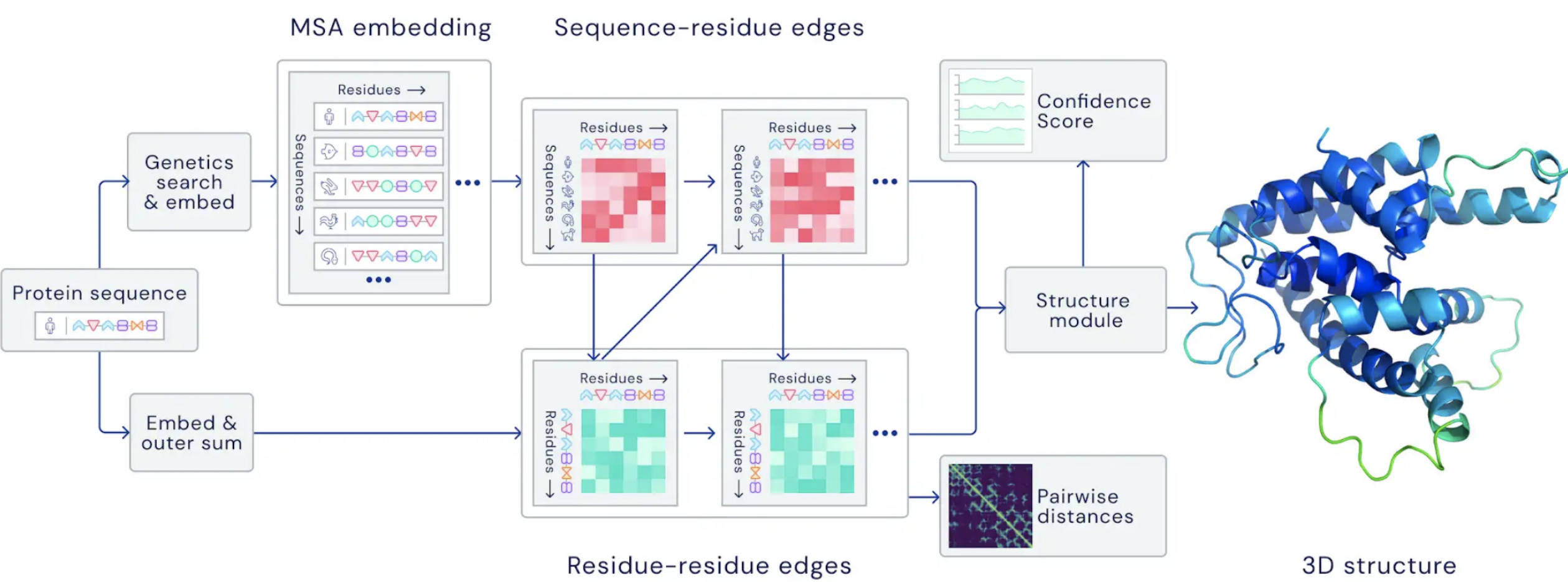


PRINCETON
UNIVERSITY

AI for Science: A Revolution?

Breakthroughs in **(Structural) Biology**

Protein Structure Prediction



AlphaFold2, Jumper et al. Nature 2021

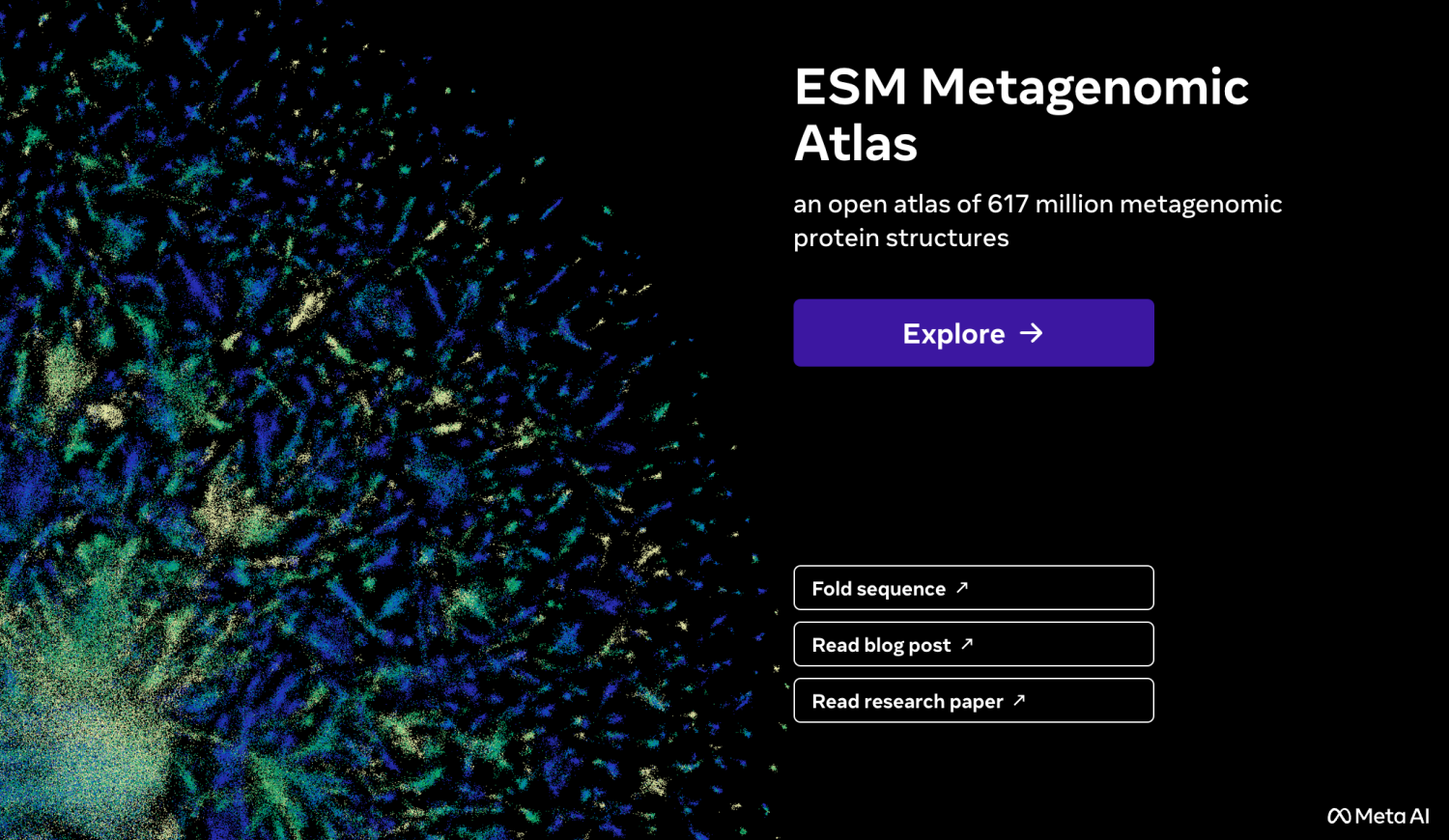
Protein Design



Chroma, Ingraham et al, Nature 2023



AlphaFold3, Nature 2024



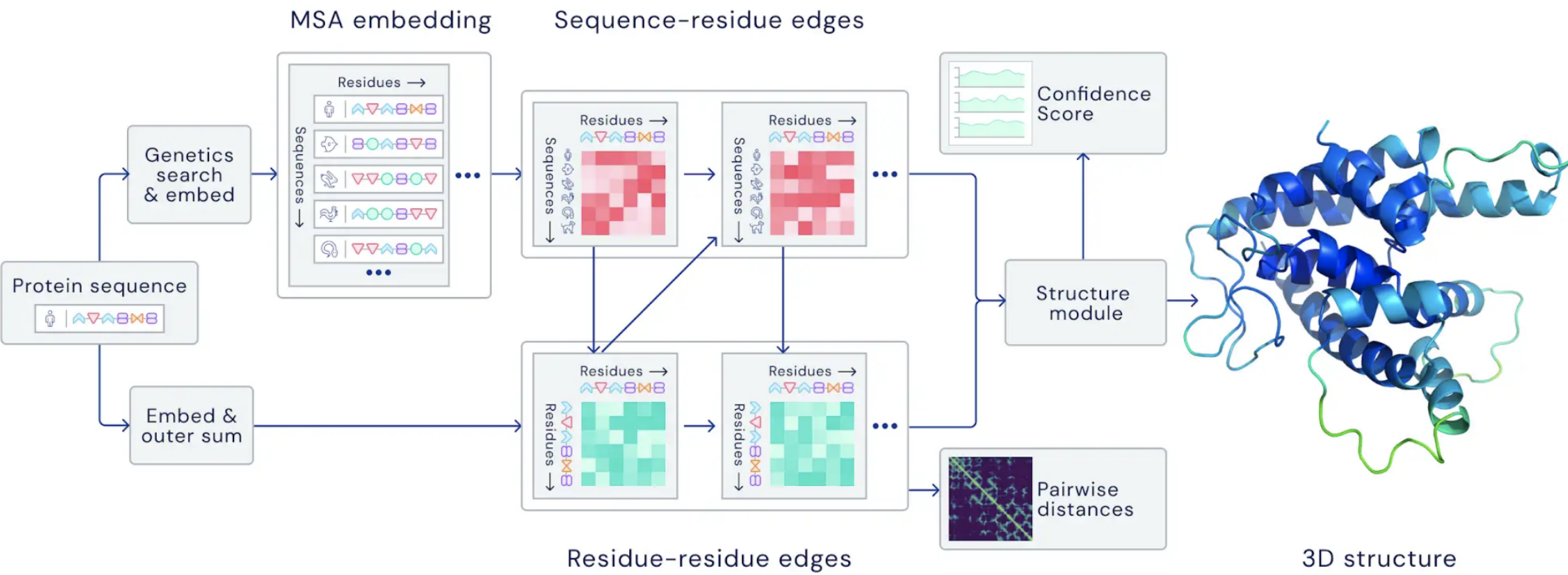
ESMfold, Lin et al, Science 2023

Also:
rosettafold,
omegafold,
librefold,
openfold,
dmpfold,
helixfold...

AI for Science: A Revolution?

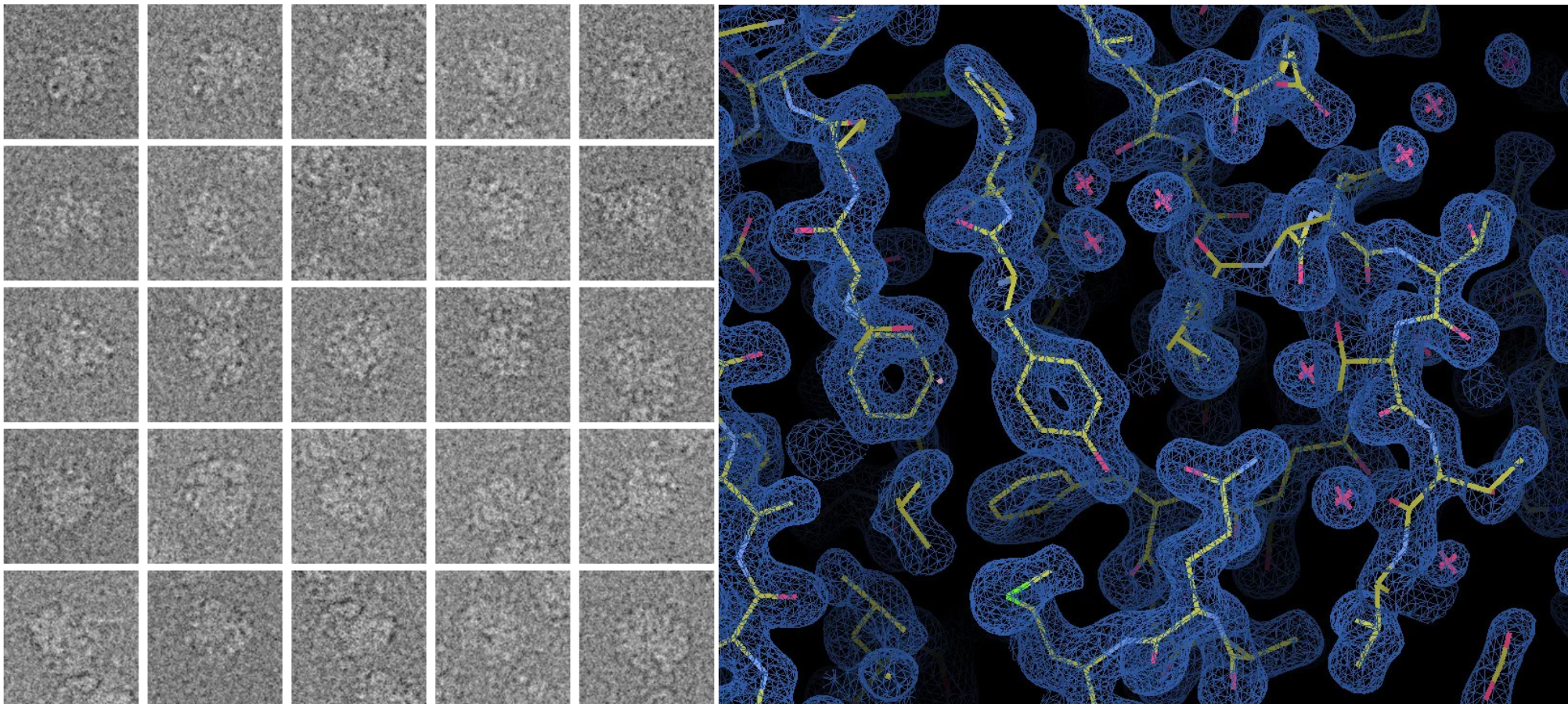
Breakthroughs in **(Structural) Biology**

Protein Structure Prediction

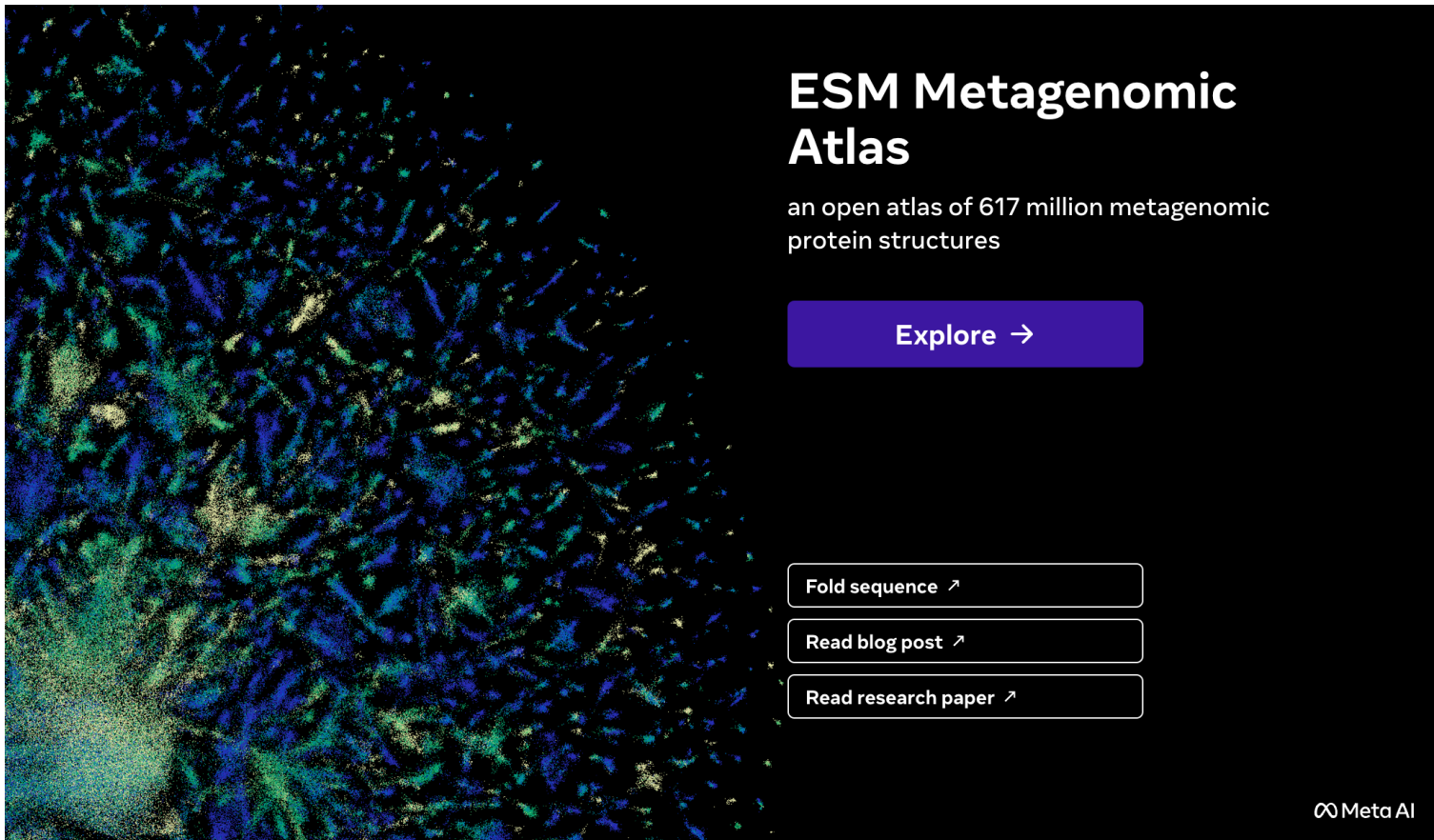


AlphaFold2, Jumper et al. Nature 2021

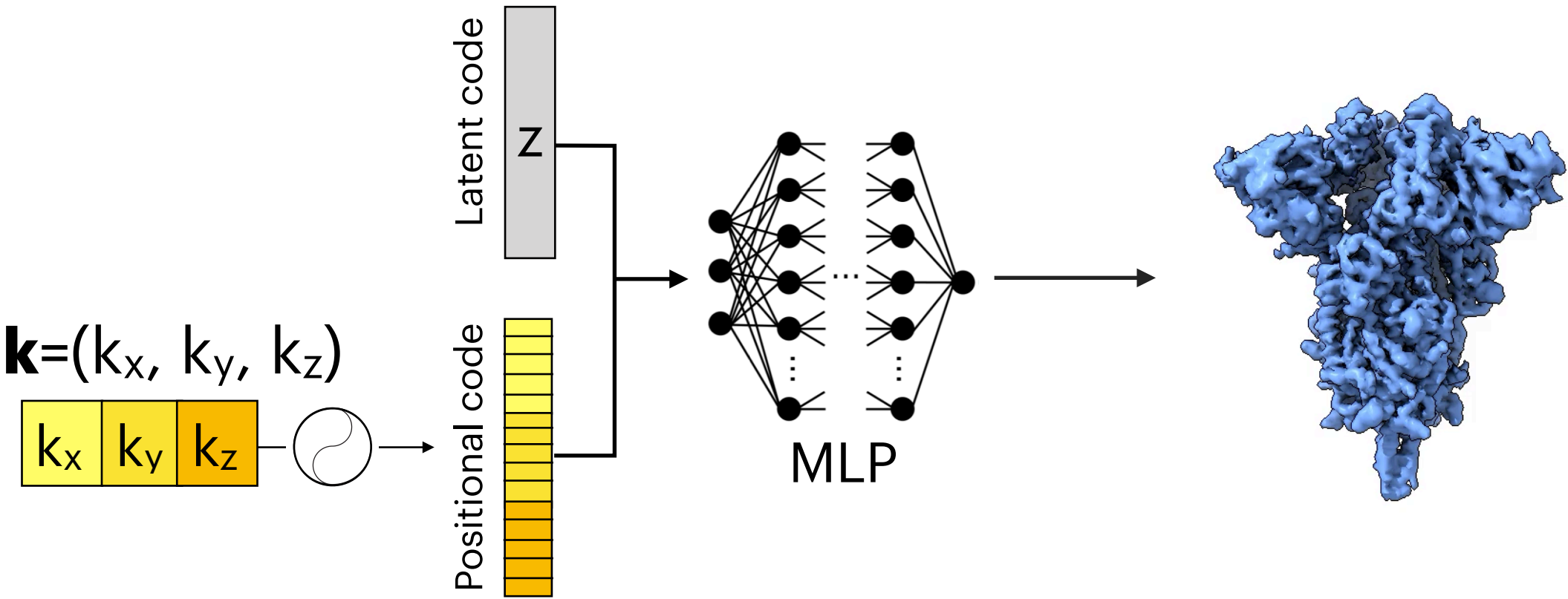
Protein Structure Determination



AlphaFold3, Nature 2024



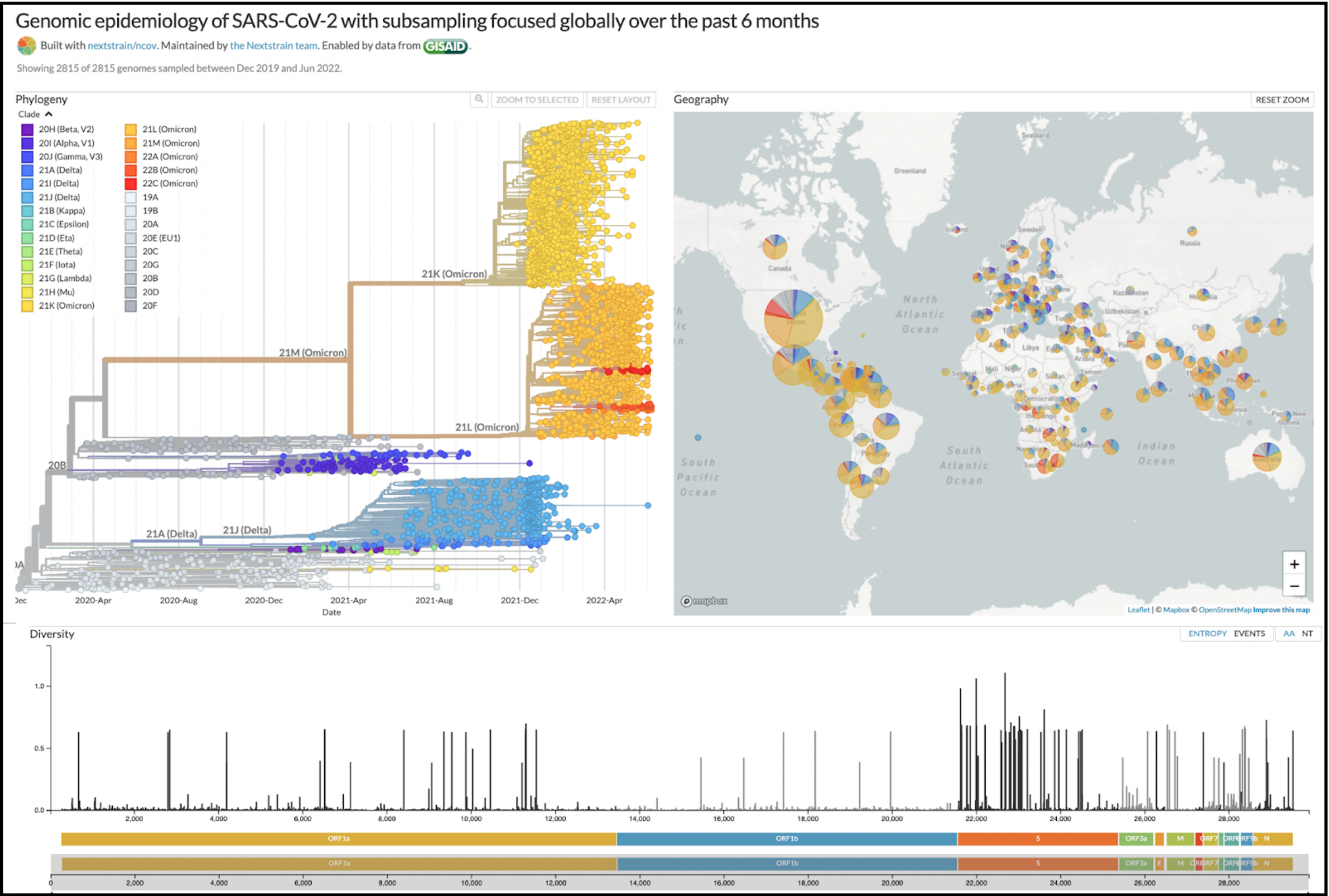
ESMfold, Lin et al, Science 2023



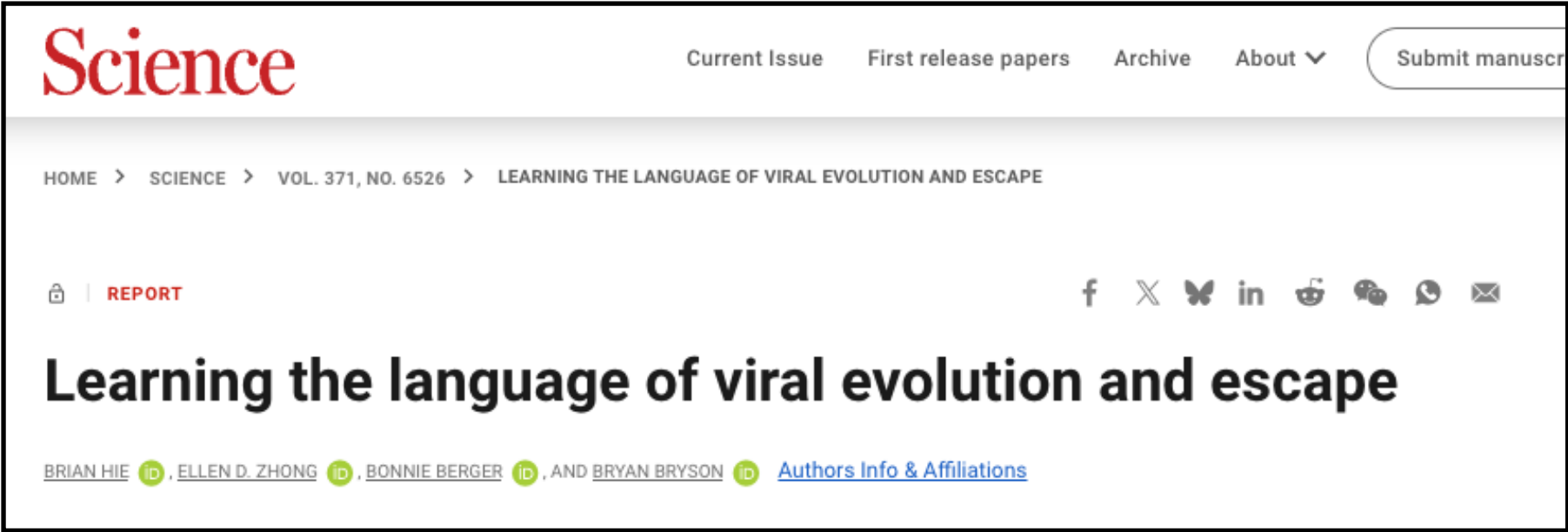
CryoDRGN. Zhong et al. ICLR 2020; Nature Methods 2021

AI for Science: A Revolution?

Modeling Biological Sequences



Source: NextStrain



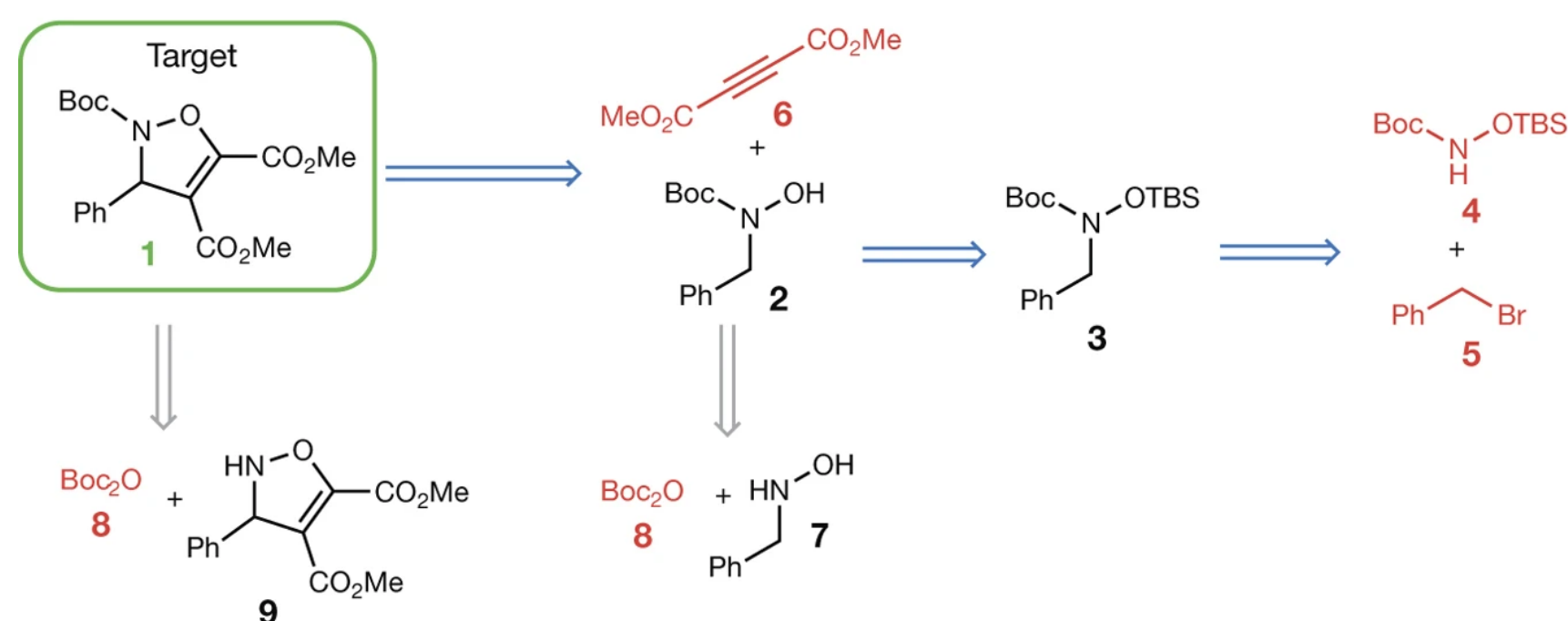
AI for Science: A Revolution?

Chemistry: AI as a scientist/agent

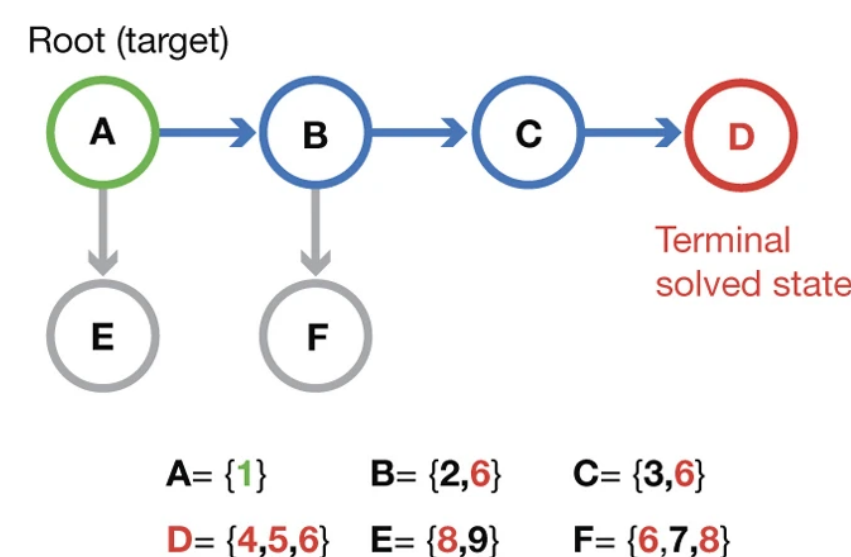
Reaction and retrosynthesis pathways

Autonomous Labs

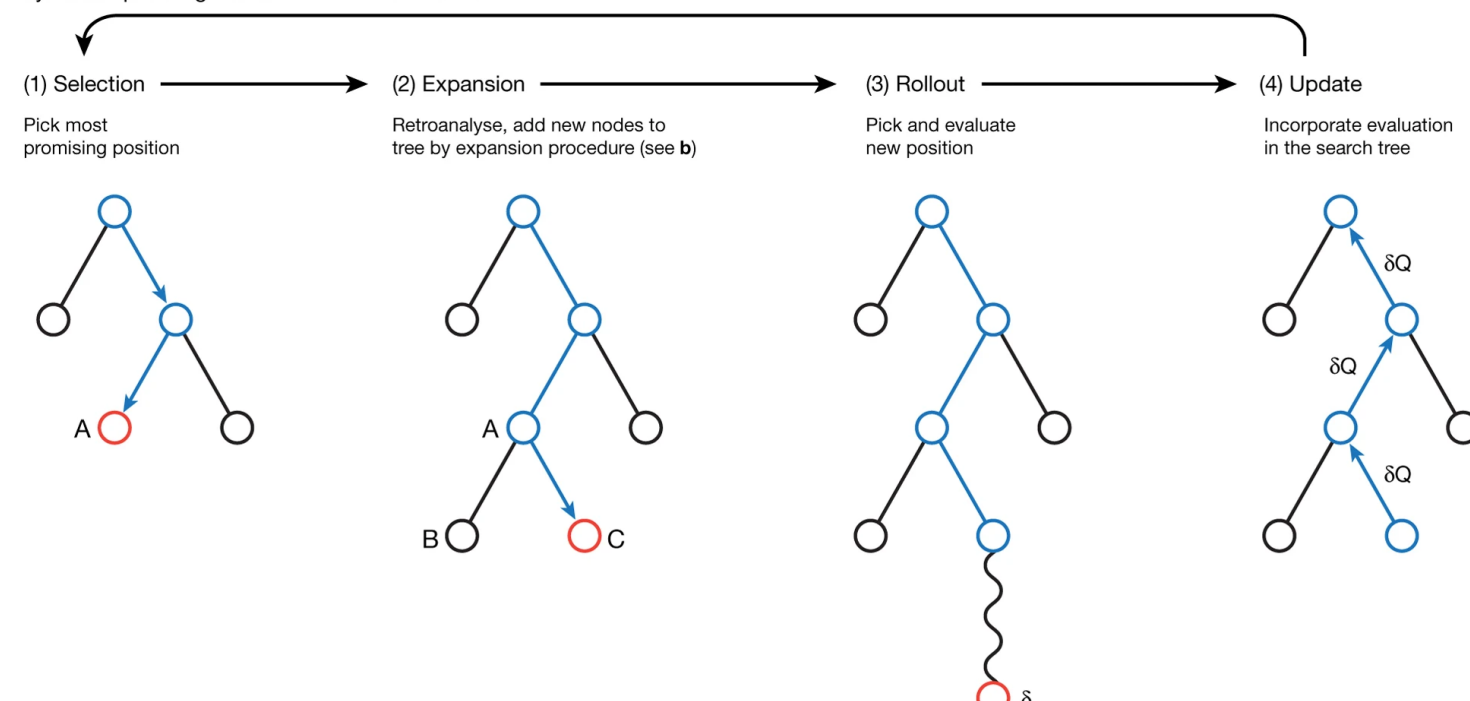
a Chemical representation of the synthesis plan



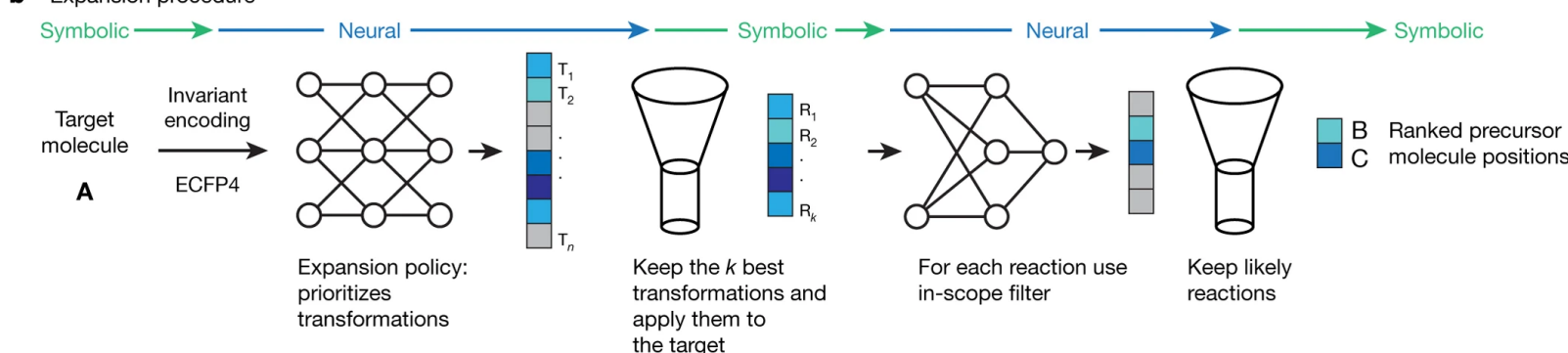
b Search tree representation



a Synthesis planning with Monte Carlo tree search



b Expansion procedure



nature [View a](#)

[Explore content](#) [About the journal](#) [Publish with us](#)

[nature](#) > [articles](#) > [article](#)

Article | Published: 29 March 2018

Planning chemical syntheses with deep neural networks and symbolic AI

[Marwin H. S. Segler](#) [Mike Preuss](#) & [Mark P. Waller](#) [✉](#)

[Nature](#) 555, 604–610 (2018) | [Cite this article](#)

[Save article](#)

106k Accesses | 1755 Citations | 815 Altmetric | [Metrics](#)

nature [View a](#)

[Explore content](#) [About the journal](#) [Publish with us](#)

[nature](#) > [articles](#) > [article](#)

Article | [Open access](#) | Published: 20 December 2023

Autonomous chemical research with large language models

[Daniil A. Boiko](#), [Robert MacKnight](#), [Ben Kline](#) & [Gabe Gomes](#) [✉](#)

[Nature](#) 624, 570–578 (2023) | [Cite this article](#)

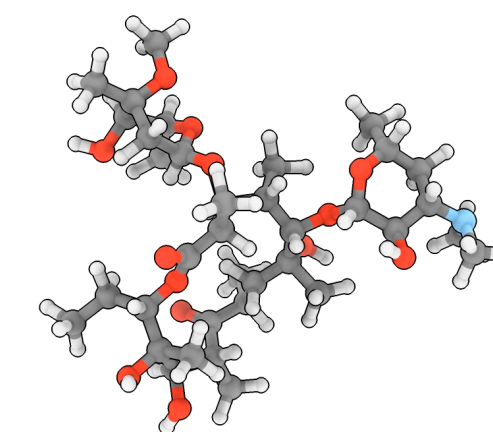
[Save article](#)

281k Accesses | 1224 Citations | 996 Altmetric | [Metrics](#)

Abstract

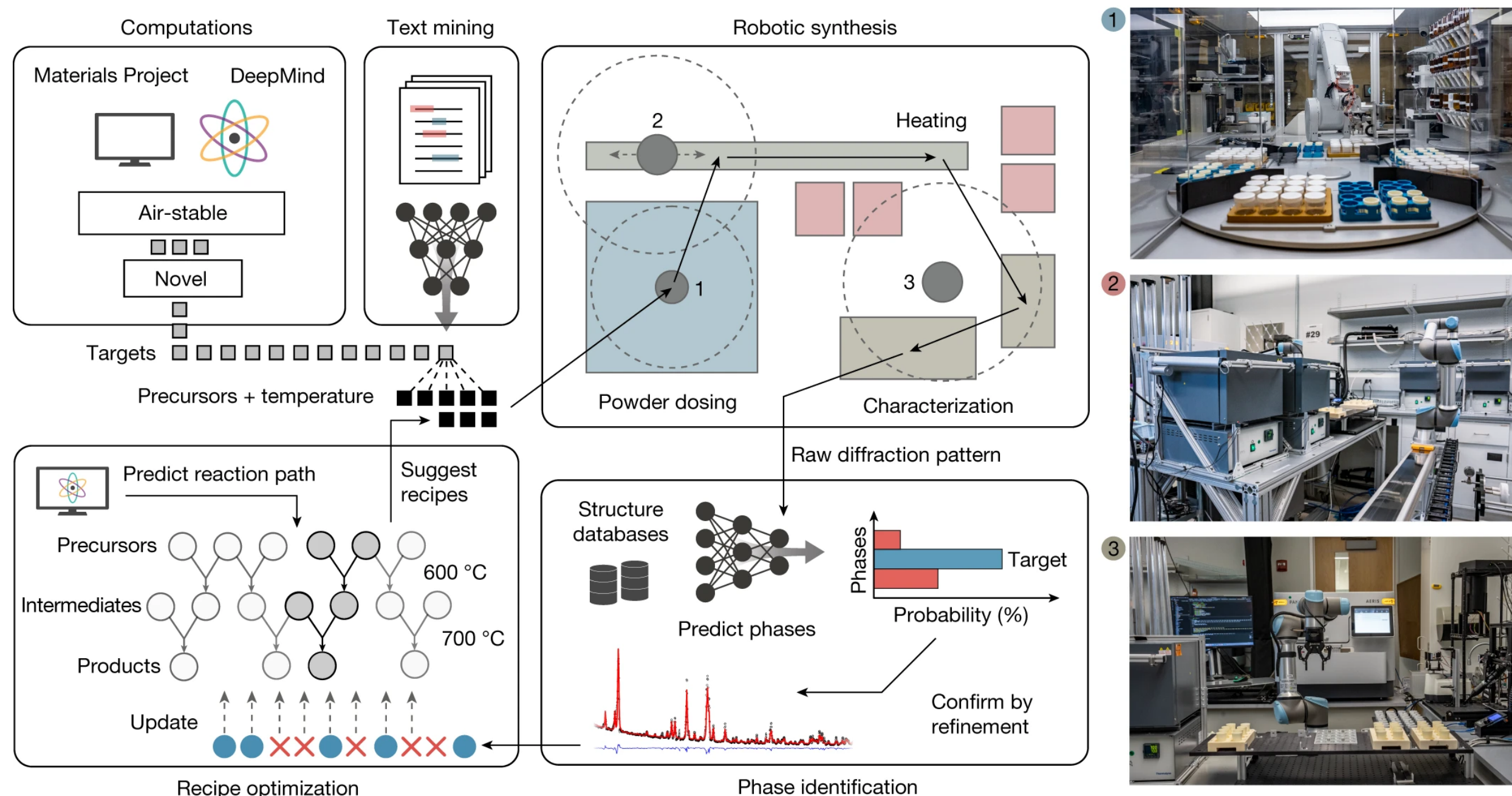
Transformer-based large language models are making significant strides in various fields, such as natural language processing^{1,2,3,4,5}, biology^{6,7}, chemistry^{8,9,10} and computer programming^{11,12}. Here, we show the development and capabilities of Coscientist, an artificial intelligence system

- Drug Discovery
- Structure Determination



AI for Science: A Revolution?

Material Science: AI as a discoverer + experimentalist



- AI can be used to search chemical space
- Autonomous labs to synthesize
- Are the materials truly novel?

nature [View all](#)

[Explore content](#) [About the journal](#) [Publish with us](#)

[nature](#) > [articles](#) > [article](#)

Article | [Open access](#) | Published: 29 November 2023

An autonomous laboratory for the accelerated synthesis of inorganic materials

[Nathan J. Szymanski](#), [Bernardus Rendy](#), [Yuxing Fei](#), [Rishi E. Kumar](#), [Tanjin He](#), [David Milsted](#), [Matthew J. McDermott](#), [Max Gallant](#), [Ekin Dogus Cubuk](#), [Amil Merchant](#), [Haegyeom Kim](#), [Anubhav Jain](#), [Christopher J. Bartel](#), [Kristin Persson](#), [Yan Zeng](#) & [Gerbrand Ceder](#)

Nature **624**, 86–91 (2023) | [Cite this article](#)

[Save article](#)

290k Accesses | 1075 Citations | 1186 Altmetric | [Metrics](#)

An [Author Correction](#) to this article was published on 19 January 2026

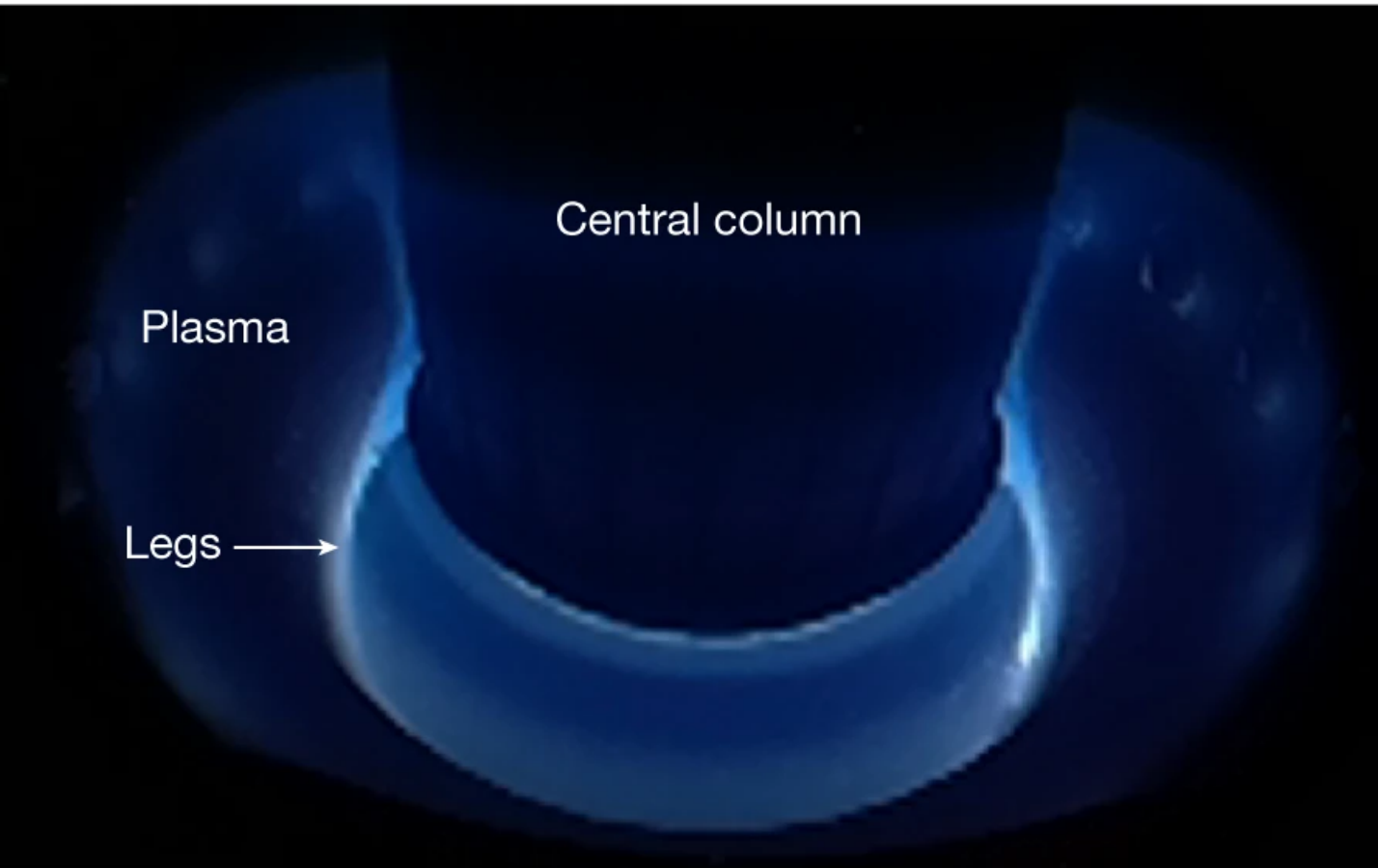
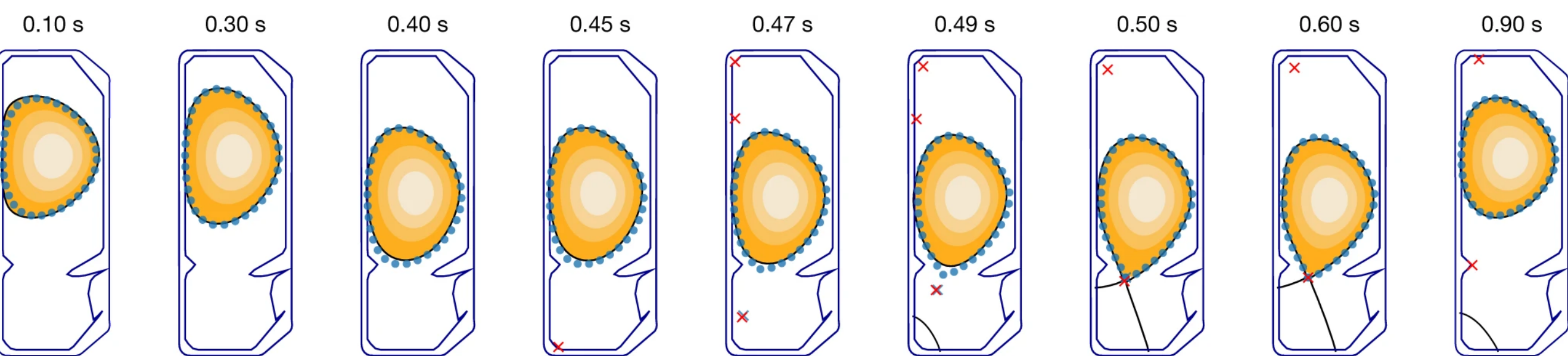
This article has been [updated](#)

Abstract

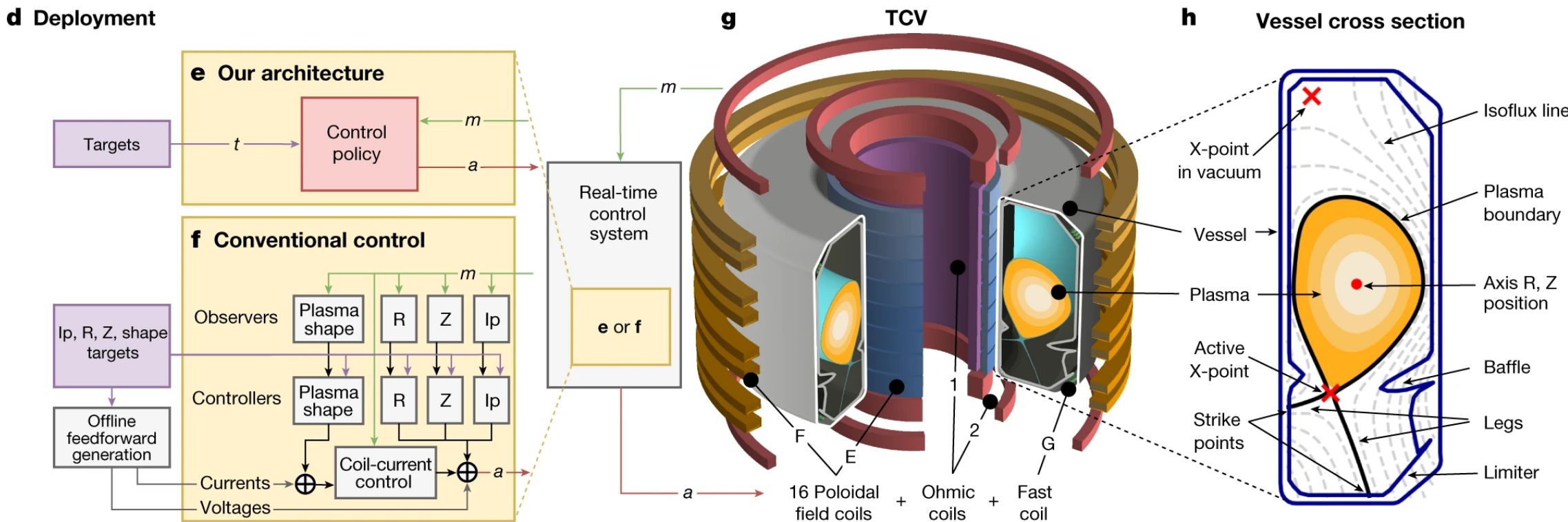
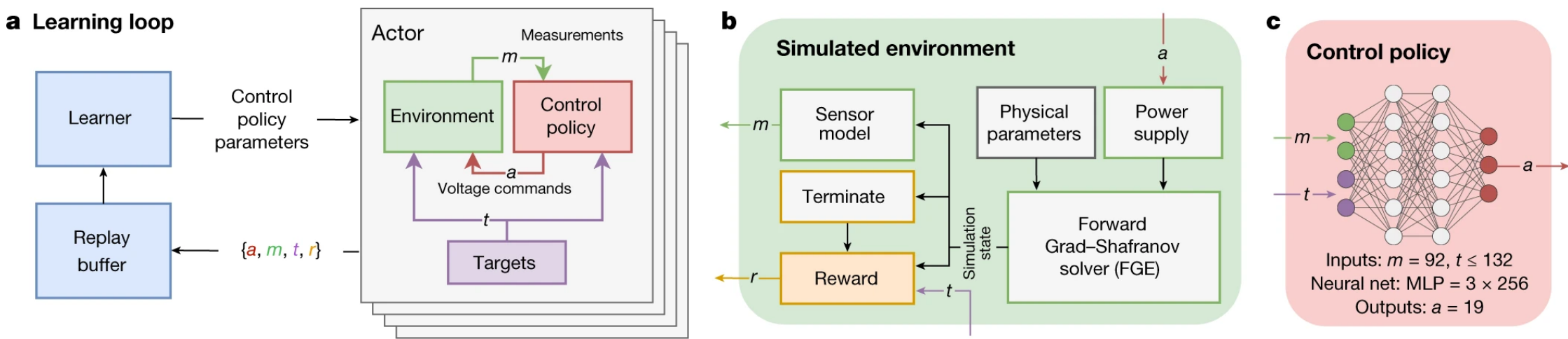
To close the gap between the rates of computational screening and experimental realization of novel materials^{1,2}, we introduce the A-Lab, an autonomous laboratory for the solid-state synthesis of inorganic powders. This platform uses computations, historical data from the literature, machine learning (ML) and active learning to plan and interpret the outcomes of experiments performed using robotics. Over 17 days of continuous operation, the A-Lab realized 36 compounds from a set of 57 targets including a variety of oxides and phosphates that were identified using large-scale ab initio phase-stability data from the Materials Project and Google DeepMind. Synthesis recipes were proposed by natural-language models trained on the literature and optimized using an active-learning approach grounded in thermodynamics. Analysis of the failed syntheses provides direct and actionable suggestions to improve current techniques for

AI for Science: A Revolution?

Physics: AI learns to **control** a fusion reactor



Inside view at 0.6 s



nature

[Explore content](#) [About the journal](#) [Publish with us](#)

[nature](#) > [articles](#) > [article](#)

Article | [Open access](#) | Published: 16 February 2022

Magnetic control of tokamak plasmas through deep reinforcement learning

[Jonas Degraeve](#), [Federico Felici](#) , [Jonas Buchli](#) , [Michael Neunert](#), [Brendan Tracey](#) , [Francesco Carpanese](#), [Timo Ewalds](#), [Roland Hafner](#), [Abbas Abdolmaleki](#), [Diego de las Casas](#), [Craig Donner](#), [Leslie Fritz](#), [Cristian Galperti](#), [Andrea Huber](#), [James Keeling](#), [Maria Tsimpoukelli](#), [Jackie Kay](#), [Antoine Merle](#), [Jean-Marc Moret](#), [Seb Noury](#), [Federico Pesamosca](#), [David Pfau](#), [Olivier Sauter](#), [Cristian Sommariva](#), [Stefano Coda](#), [Basil Duval](#), [Ambrogio Fasoli](#), [Pushmeet Kohli](#), [Koray Kavukcuoglu](#), [Demis Hassabis](#) & [Martin Riedmiller](#)  — [Show fewer authors](#)

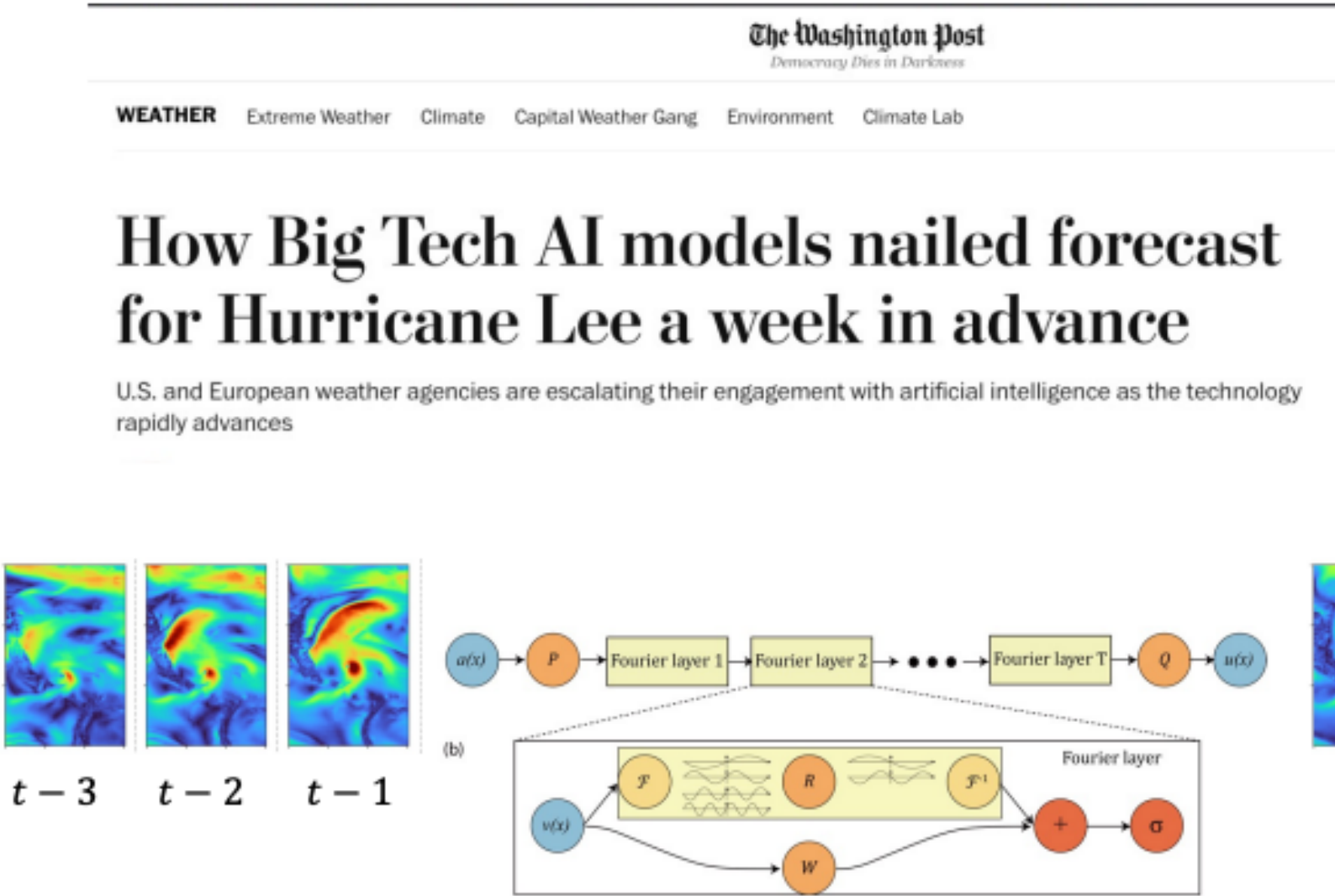
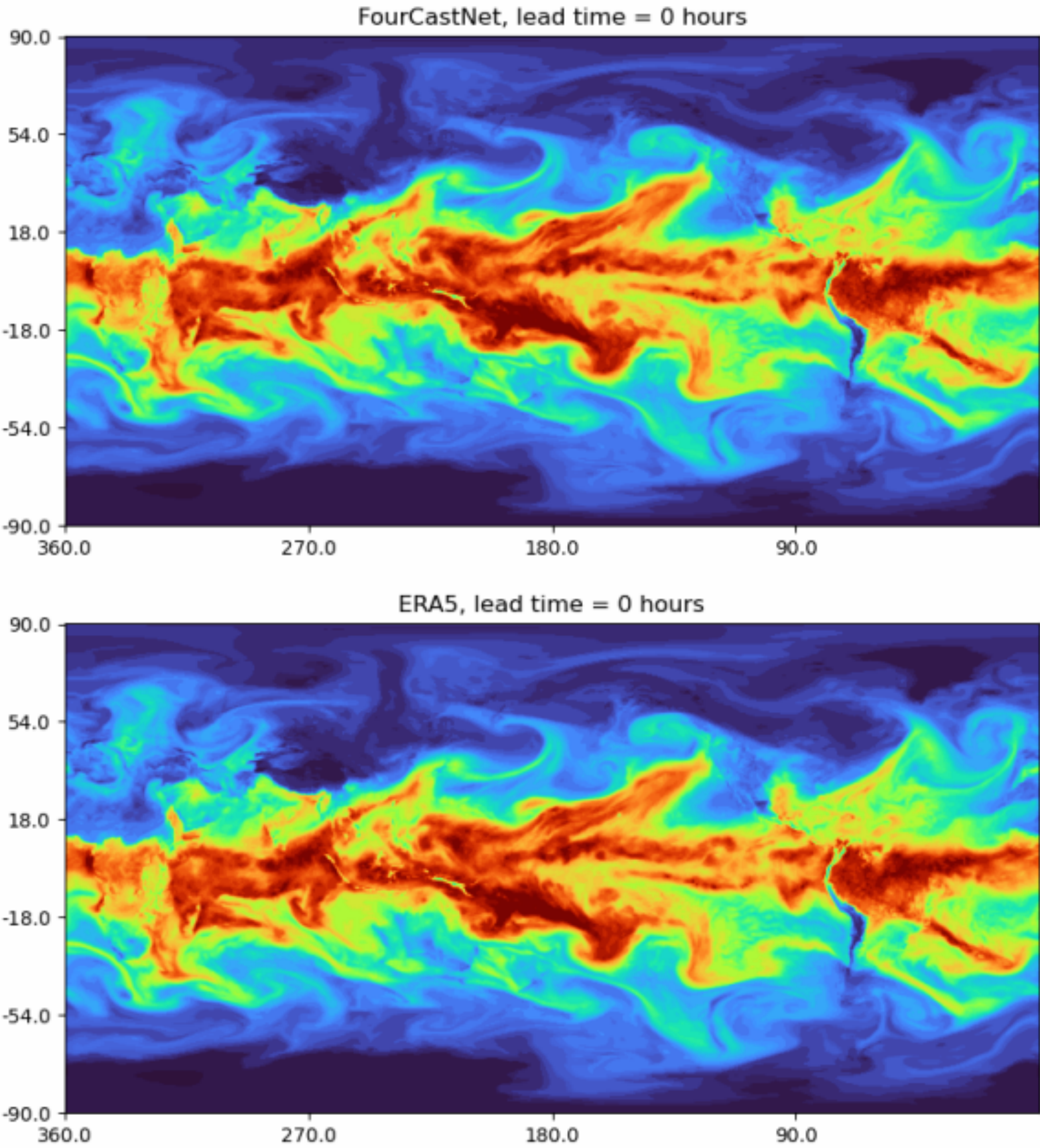
[Nature](#) **602**, 414–419 (2022) | [Cite this article](#)

 [Save article](#)

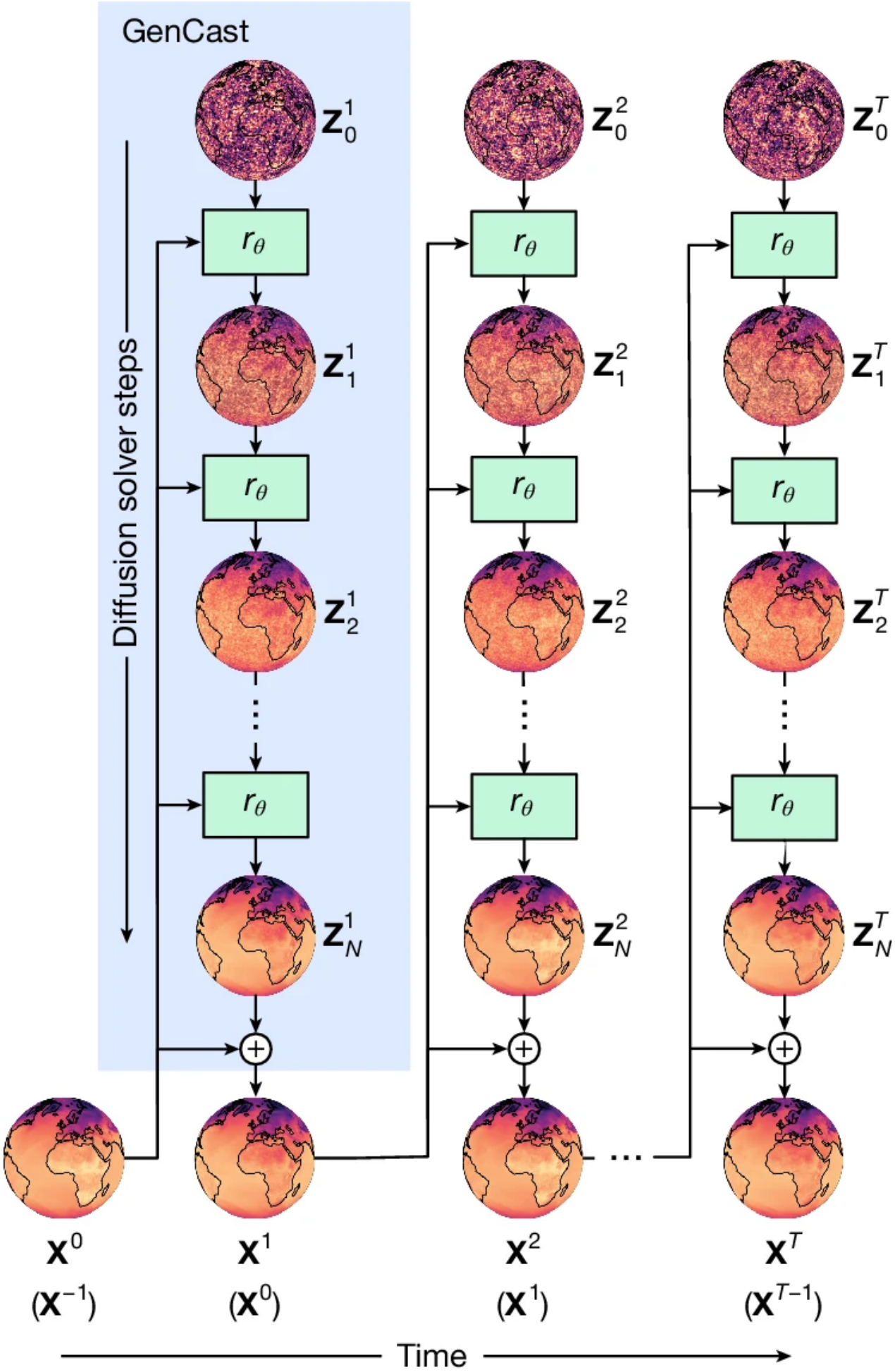
309k Accesses | **823** Citations | **2483** Altmetric | [Metrics](#)

AI for Science: A Revolution?

Weather Forecasting: AI as a simulator



Pathak et al, FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators, ArXiv (2022)



nature [View all](#)

[Explore content](#) [About the journal](#) [Publish with us](#)

nature > articles > article

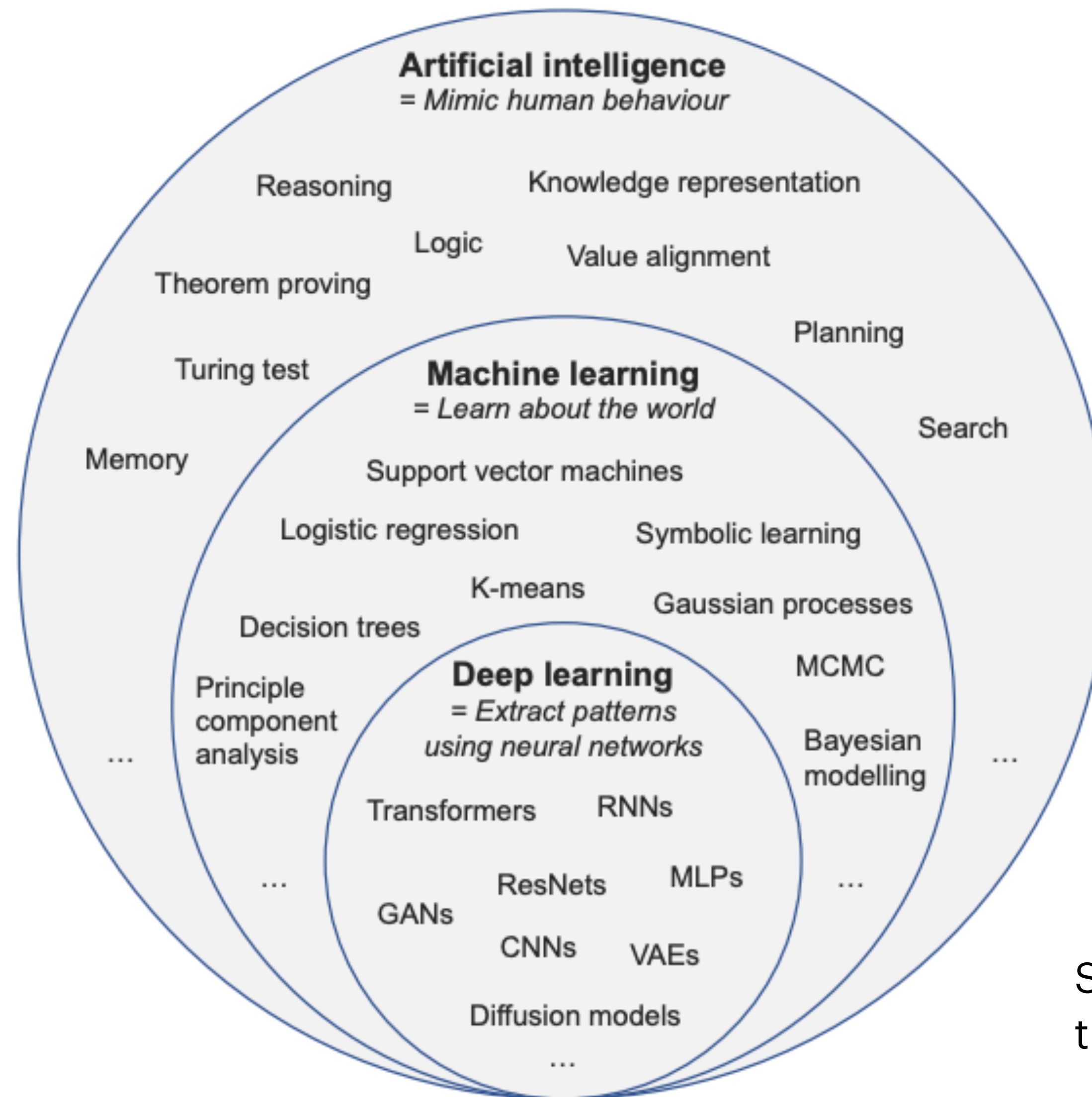
Article | [Open access](#) | Published: 04 December 2024

Probabilistic weather forecasting with machine learning

[Ilan Price](#) [Alvaro Sanchez-Gonzalez](#) [Ferran Alet](#) [Tom R. Andersson](#) [Andrew El-Kadi](#) [Dominic Masters](#) [Timo Ewalds](#) [Jacklynn Stott](#) [Shakir Mohamed](#) [Peter Battaglia](#) [Remi Lam](#) & [Matthew Willson](#)

[Nature](#) 637, 84–90 (2025) | [Cite this article](#)

Deep learning vs. machine learning vs. AI



Should this course be called “AI for the natural sciences”?

How is AI used in Science?

- Let's organize by what the model does instead of by field

Surrogates for expensive computation. Learn the output of a simulator you already trust — MLIPs standing in for DFT, structure prediction standing in for physics-based folding. The win is throughput, which changes what you can search over.

Inverse problems. Recover latent structure from indirect, noisy measurements: cryo-EM and cryo-ET, NMR, crystallography, mass spec, most of microscopy. Well-posed enough to benchmark, and the place where the prior-vs-data question actually bites.

Generative design. Sample from a distribution over molecules, sequences, or synthesis routes. Success is defined by wet-lab hit rates, so evaluation is external and slow.

Representation learning at scale. PLMs, cell atlases, sequence-to-function. Extract structure from unlabeled data, then transfer. Hardest to evaluate because the claim is about generality.

Experiment selection. Active learning, BO, self-driving labs. Most cited as a promise, thinnest evidence base.

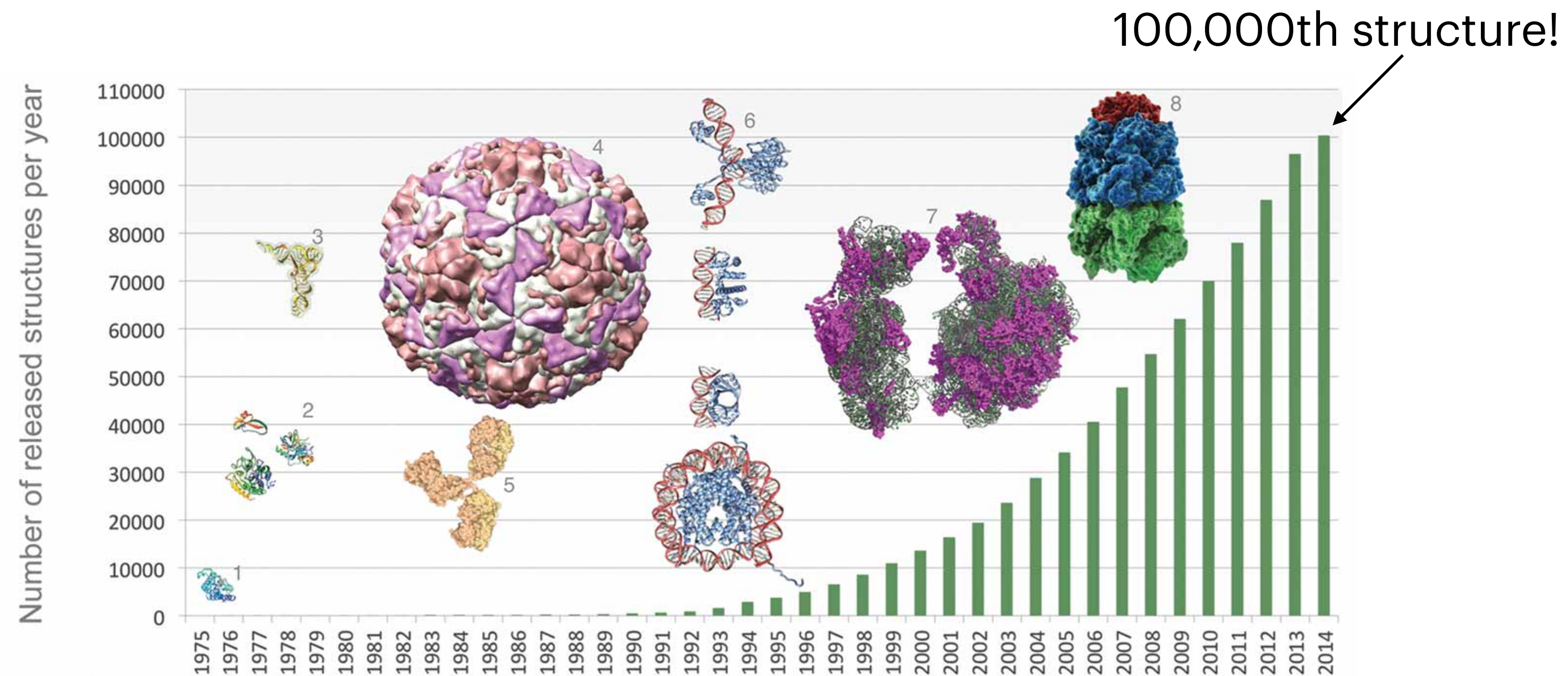
The assistant layer. Literature, code, agents. Real productivity effects, almost no measurement.

How is AI used in Science?

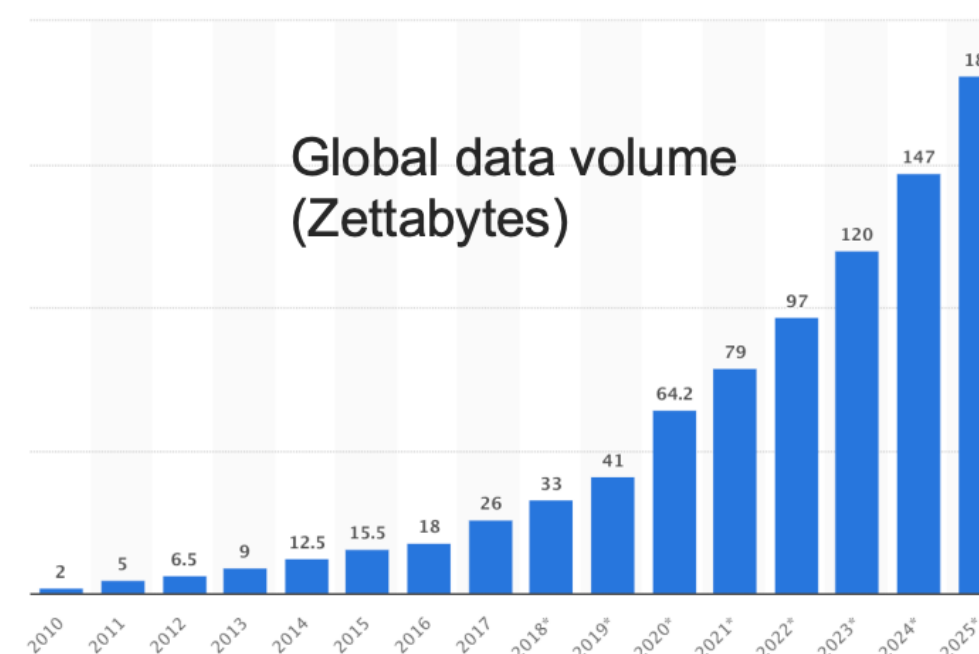
Role of AI	What AI does	Examples
1. Perception & measurement	Extracts scientific information from raw observations	Cryo-EM reconstruction, microscopy segmentation, astronomical object detection, spectroscopy interpretation
2. Prediction	Learns $x \rightarrow y$ mappings from data	Protein structure, molecular properties, weather, materials properties
3. Representation & compression	Finds low-dimensional structure in complex data	Protein embeddings, latent spaces of molecular conformations, dimensionality reduction
4. Generative modeling	Generates plausible scientific objects or states	Proteins, molecules, materials, structures, conformational ensembles
5. Simulation & surrogate modeling	Approximates expensive physical calculations	Neural potentials, learned PDE solvers, climate emulators
6. Inverse problems	Infers hidden causes/structures from observations	Structure from cryo-EM/NMR, tomography, inverse design
7. Hypothesis generation	Finds patterns and proposes explanations or candidates	Gene–disease relationships, reaction mechanisms, candidate materials
8. Experimental design	Decides what measurement/experiment to perform next	Active learning, Bayesian optimization, adaptive microscopy
9. Automated experimentation	Couples AI to instruments/robots	Self-driving chemistry labs, autonomous materials synthesis
10. Scientific reasoning	Reads literature, integrates evidence, plans analyses and experiments	Scientific agents, literature agents, coding/data-analysis agents
11. Discovery	Searches spaces too large for humans to explore manually	New proteins, drugs, materials, algorithms, mathematical conjectures

Some Challenges in AI for Science

- Learning from scientific data
 - Data scarcity / efficiency
 - Generalization
 - Data quality



The PDB



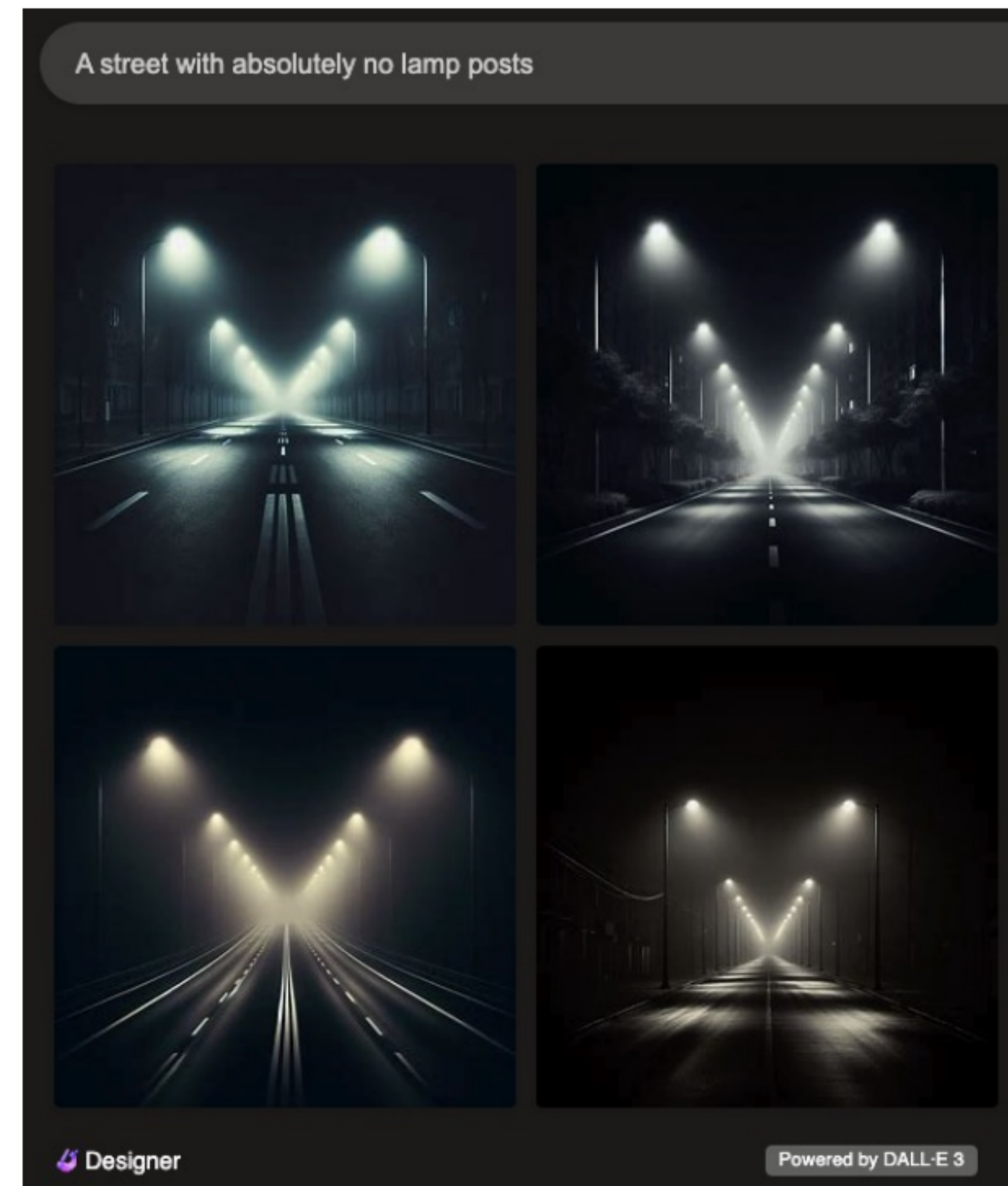
Source: Statista

IMAGENET



Some Challenges in AI for Science

- Learning from scientific data
 - Data scarcity / efficiency
 - Generalization
 - Data quality
- Scientific understanding
 - Reasoning
 - Lack of interpretability
 - Physical consistency



Source: DALL·E 3

Some Challenges in AI for Science

- Learning from scientific data
 - Data scarcity / efficiency
 - Generalization
 - Data quality
- Scientific understanding
 - Reasoning
 - Lack of interpretability
 - Physical consistency
- Evaluation and reliability
 - Data leakage
 - Benchmark and metrics
 - Uncertainty & calibration
 - Reproducibility / robustness

Leakage and the Reproducibility Crisis in ML-based Science

Sayash Kapoor¹ Arvind Narayanan¹

Abstract

The use of machine learning (ML) methods for prediction and forecasting has become widespread across the quantitative sciences. However, there are many known methodological pitfalls, including data leakage, in ML-based science. In this paper, we systematically investigate reproducibility issues in ML-based science. We show that data leakage is indeed a widespread problem and has led to severe reproducibility failures. Specifically, through a survey of literature in research communities that adopted ML methods, we find 17 fields where errors have been found, collectively affecting 329 papers and in some cases leading to wildly overoptimistic conclusions. Based on our survey, we present a fine-grained taxonomy of 8 types of leakage that range from textbook errors to open research problems.

Rest of this lecture

- Course overview and logistics
- The scientific process
- The taxonomy of “AI for Science”
- Discussion

Course Introduction

Welcome!

- Introductions:
 - What is your year / background / research interests?
 - What do you want to learn from this class?
 - Fun fact?

Logistics

- Website: <https://www.cs.princeton.edu/courses/archive/fall26/cos597Y/>
- Instructor: Ellen Zhong (she/her)
- Office hours: Tuesdays 3:00-4:00p, CS 314, or by appointment
- Class meetings:
 - Tentative: Wednesdays 3-5p only (Will poll via Google Form for availability)
 - Attendance is required — contact me in advance if there are extenuating circumstances
- We will mostly use slack for communication and to share helpful resources or related papers, or #random! I will create a slack channel after this meeting and add everyone.

Learning Goals

- Understand how scientific data and domain structure shape machine learning methods
- Become familiar with major applications of AI across the natural sciences
- Understand the strengths, limitations, and appropriate uses of AI for scientific problems
- Learn how to formulate research problems and communicate scientific ideas
- Gain exposure to academic and industry perspectives on AI for science

Format

- There are two components of this class:
 - In-class presentations and discussions on assigned papers
 - Final project
- See syllabus for more info: <https://www.cs.princeton.edu/courses/archive/fall26/cos597Y/>

Grading

- Participation (30%)
- Presentation (30%)
- Final project (40%)
- Participation
 - One of the primary ways of learning and engaging with the material
 - Examples: Posing or answering questions, explaining background material, sharing reflections
 - Pre-lecture questions
 - Post-lecture feedback
 - Class attendance is mandatory. Please reach out to course staff (Ellen + Ziyu) if you have a conflict for a makeup assignment

Overview of Topics

- Before fall break
 - AlphaFold :)
 - Intro to deep learning and generative modeling
 - Transformers and scaling laws
 - Diffusion & flow matching
 - Inverse problems
 - Scientific foundation models

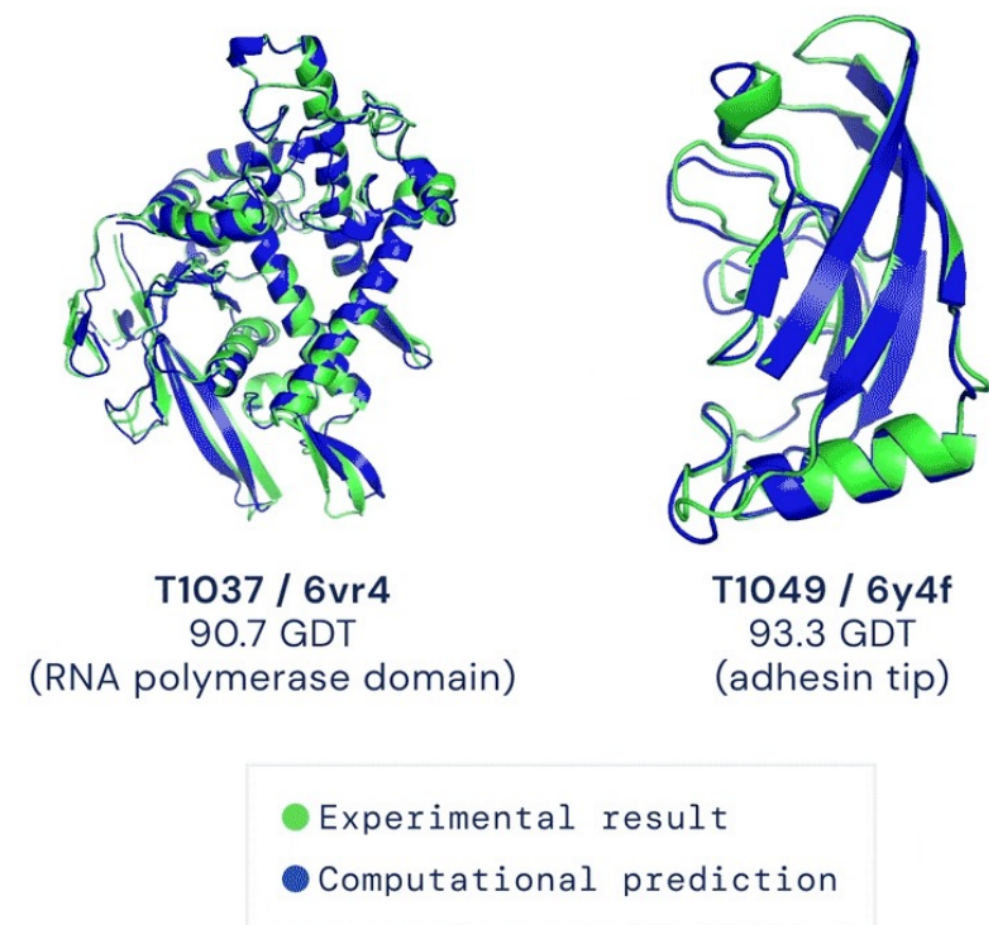
Overview of Topics

- After fall break we will focus on applications:
 - Biological language modeling
 - Geometric DL, AI for chemistry
 - AI for physics & learning dynamics
 - AI agents for science
 - Protein design, Bayesian optimization, active learning
 - Interpreting scientific foundation models
- The schedule is still tentative depending on MW vs W scheduling and enrollment.

What is AI for Science?

Multiple perspectives on this “field”

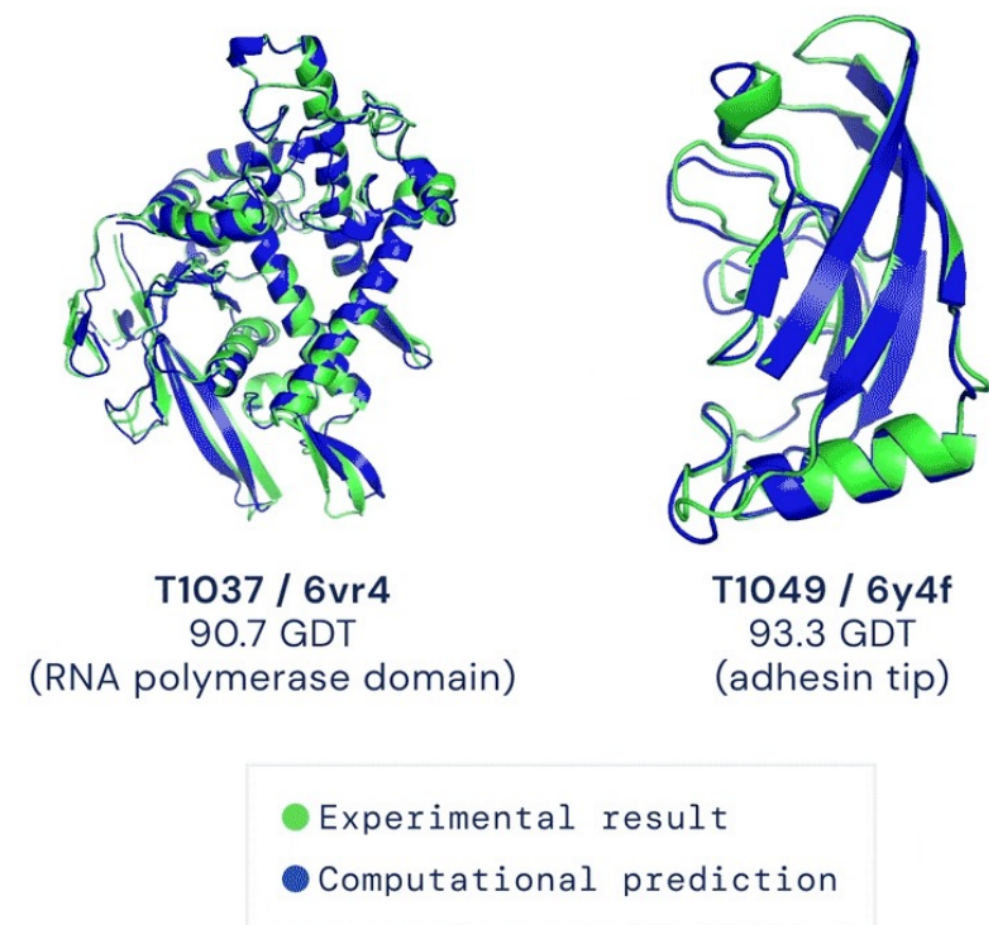
AI for **Science**



- Using AI to accelerate and/or expand the capabilities of scientific discovery
- Science is the goal; AI is the tool
- Topics: Prediction · inference · simulation · design

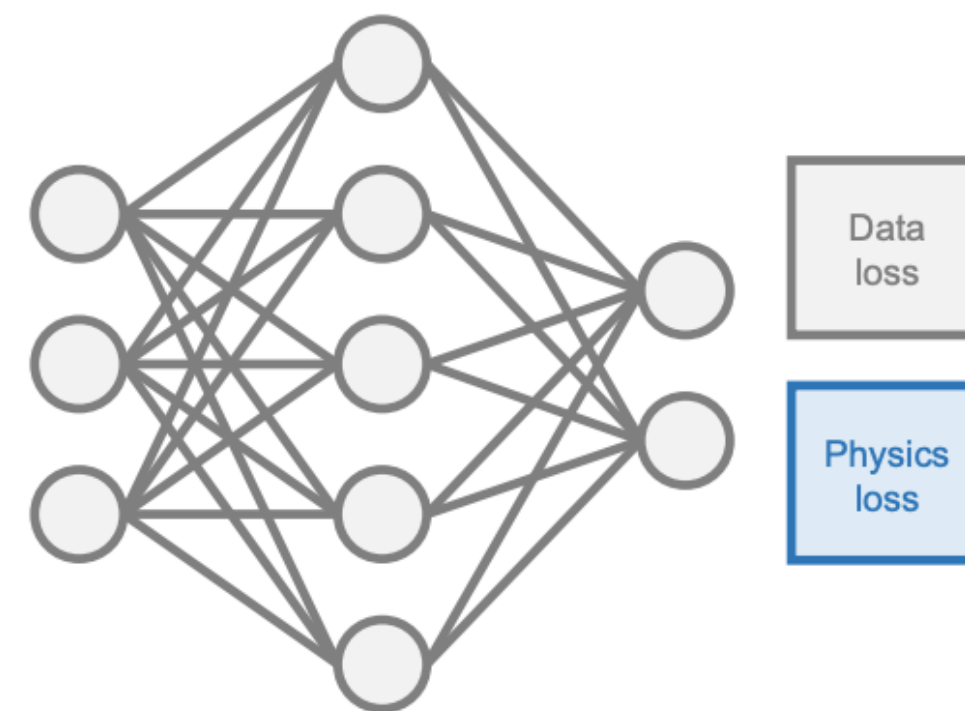
Multiple perspectives on this “field”

AI for **Science**



- Using AI to accelerate and/or expand the capabilities of scientific discovery
- Science is the goal; AI is the tool
- Topics: Prediction · inference · simulation · design

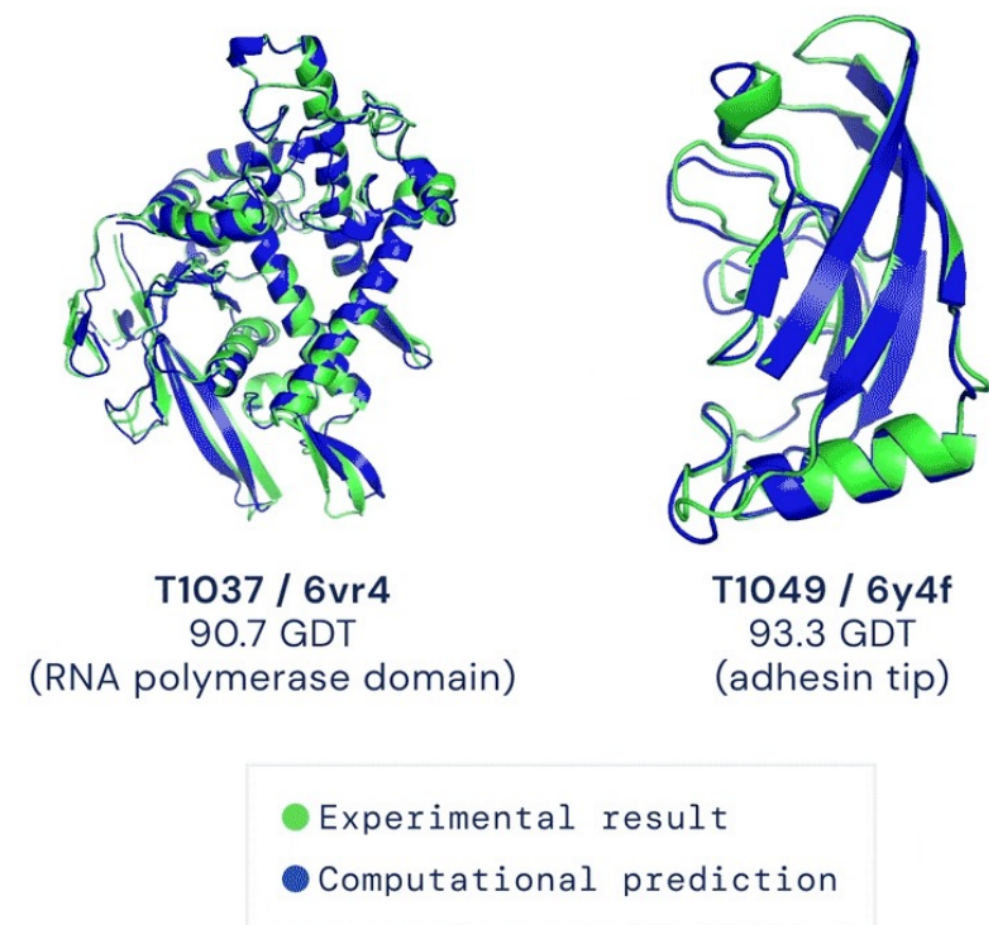
AI for Science



- Developing AI methods that incorporate scientific structure, knowledge, or constraints.
- Science informs the AI
- Topics: Physical priors · symmetries · constraints · interpretability

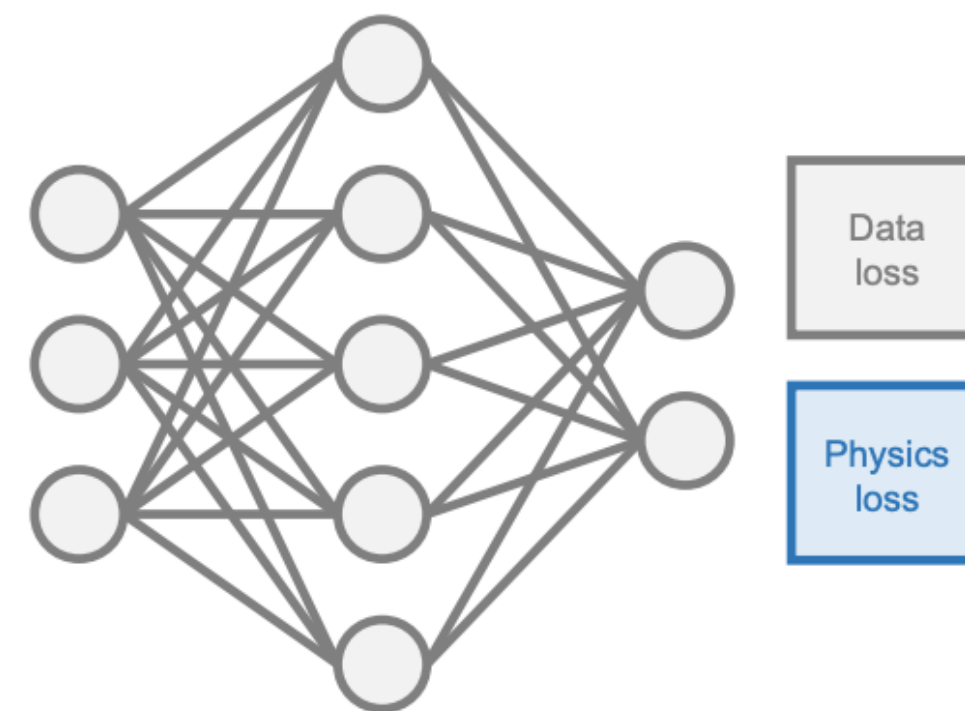
Multiple perspectives on this “field”

AI for **Science**



- Using AI to accelerate and/or expand the capabilities of scientific discovery
- Science is the goal; AI is the tool
- Topics: Prediction · inference · simulation · design

AI for Science



- Developing AI methods that incorporate scientific structure, knowledge, or constraints.
- Science informs the AI
- Topics: Physical priors · symmetries · constraints · interpretability

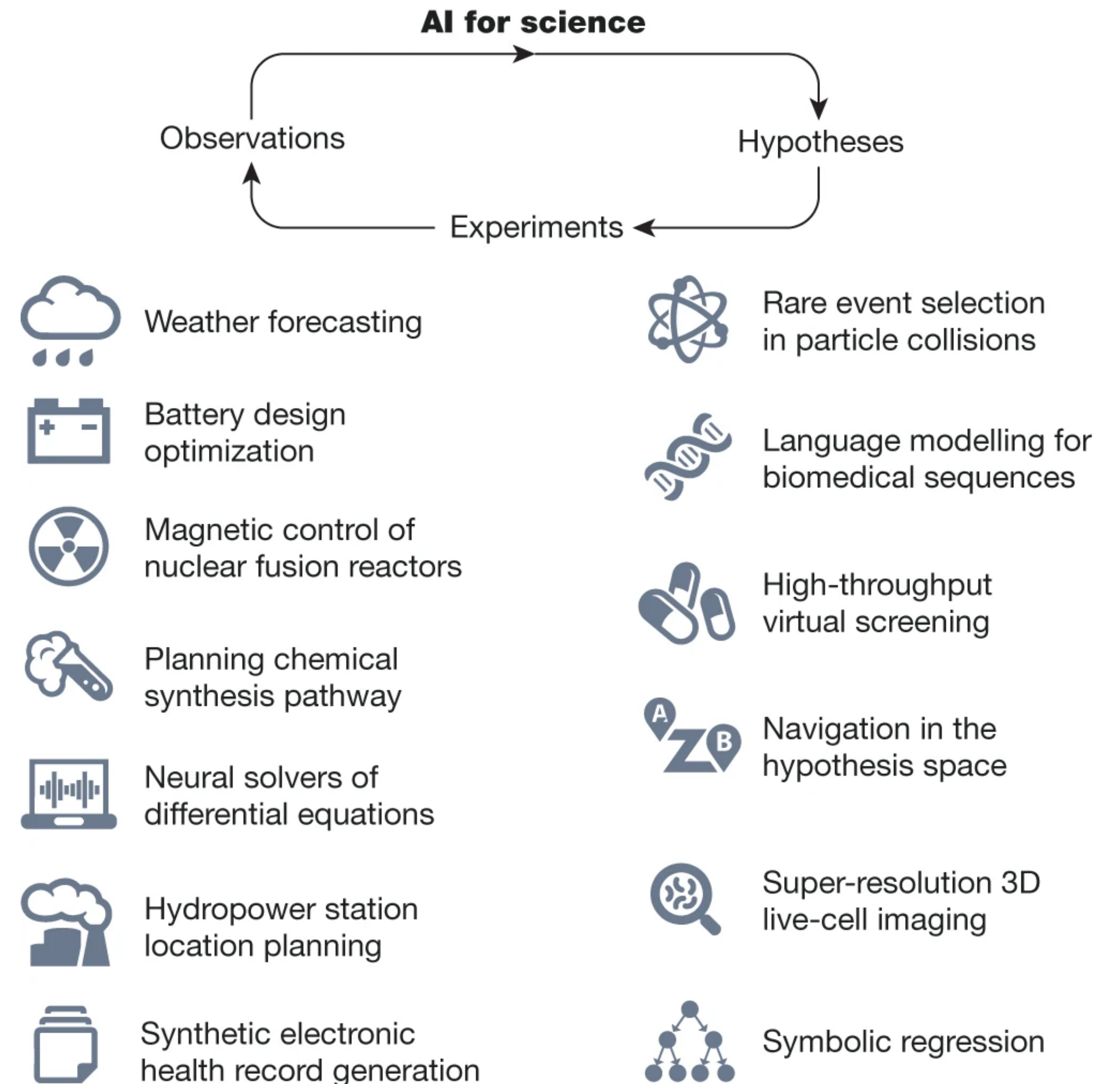
AI Scientists



- Can AI perform the process of science?
- Topics: Hypothesis generation · experiment design · autonomous labs · closed-loop discovery

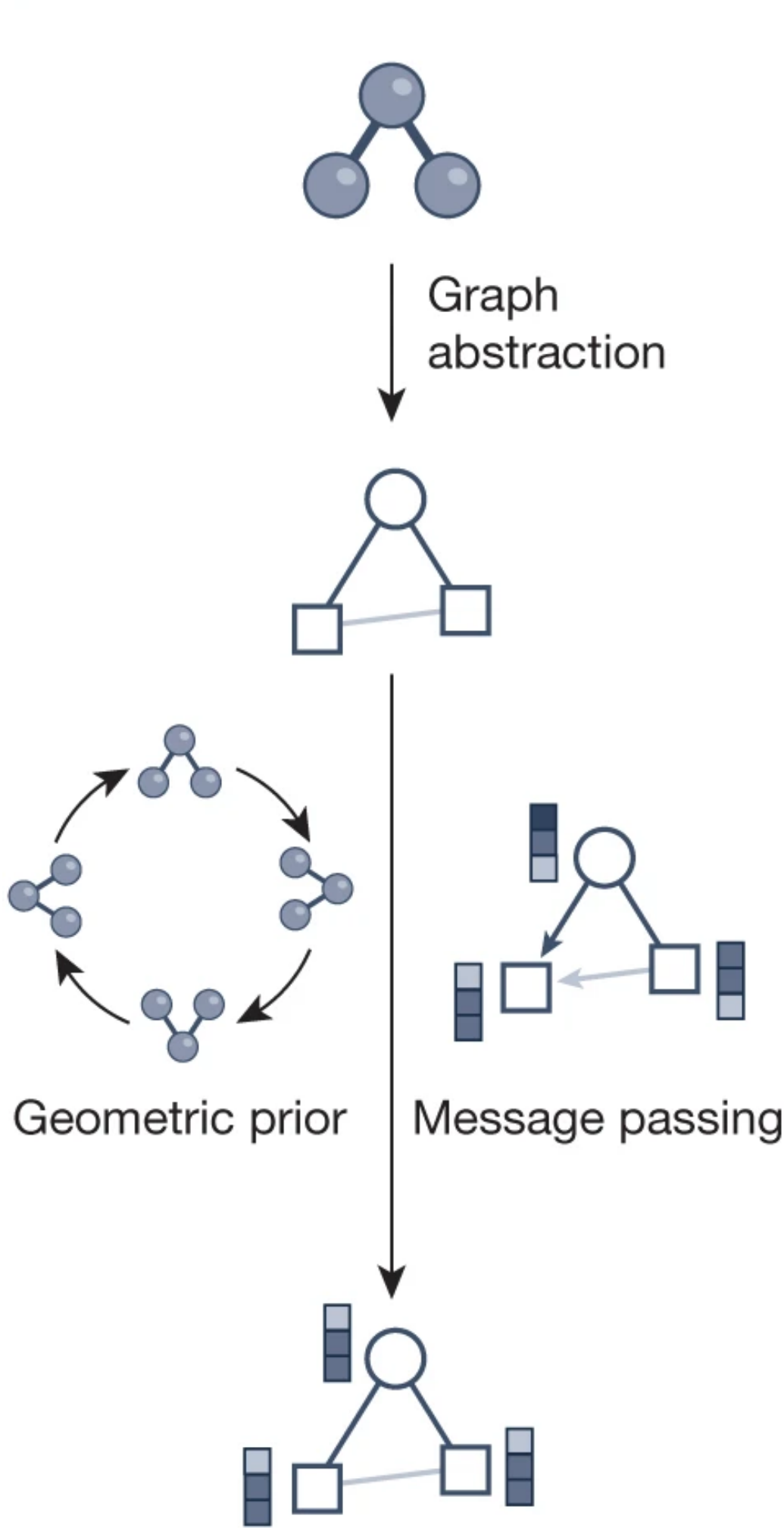
The Scientific Process

- The scientific discovery process / cycle:
 - observations → patterns → hypotheses → predictions → experiments → observations
- Where is AI (traditionally) applied in this cycle?
- Where do typical ML tasks fit in the scientific process?
 - Prediction
 - Inference
 - Generation
 - Design

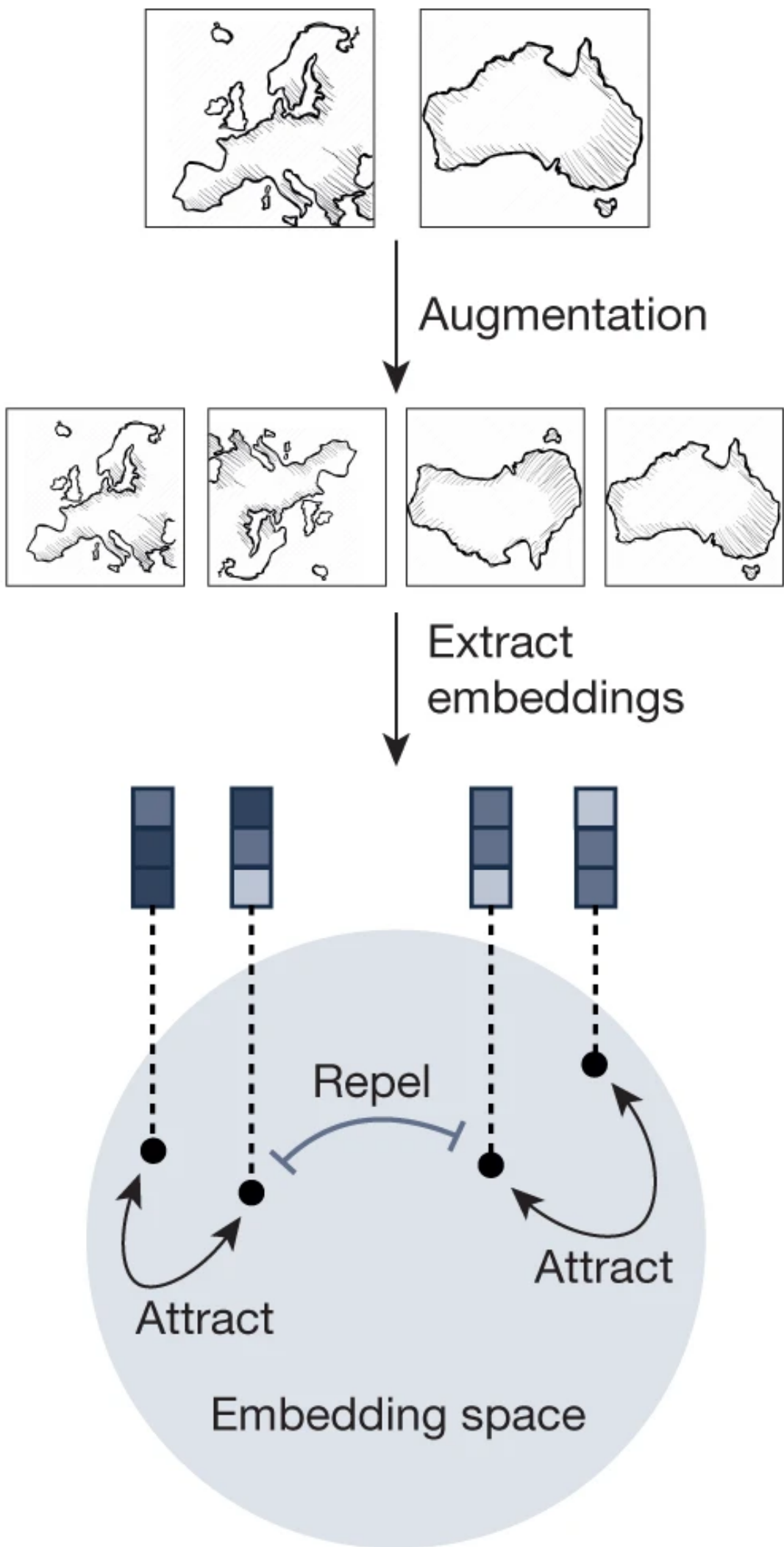


Example: Learning meaningful representations of scientific data

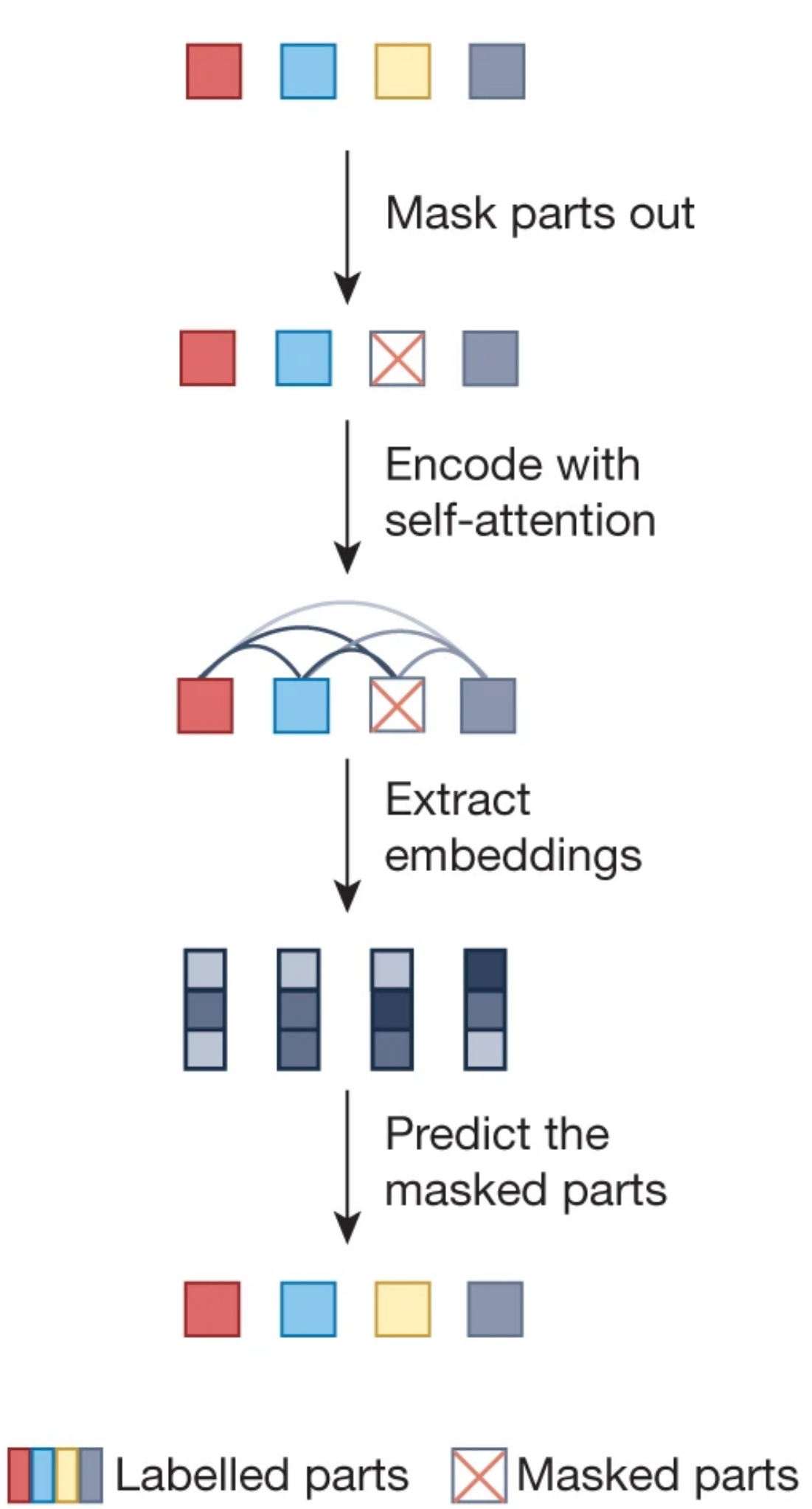
a Geometric deep learning



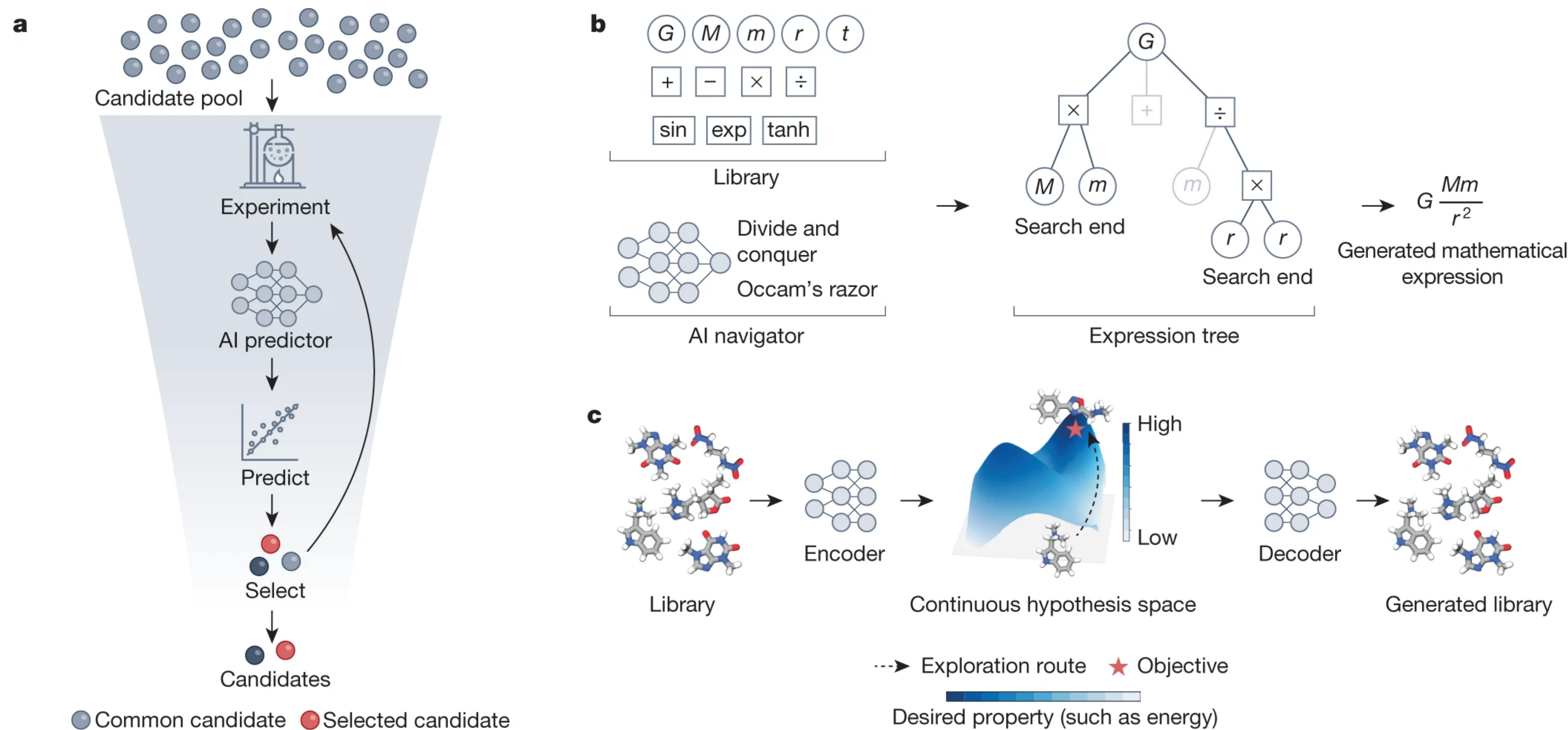
b Self-supervised learning



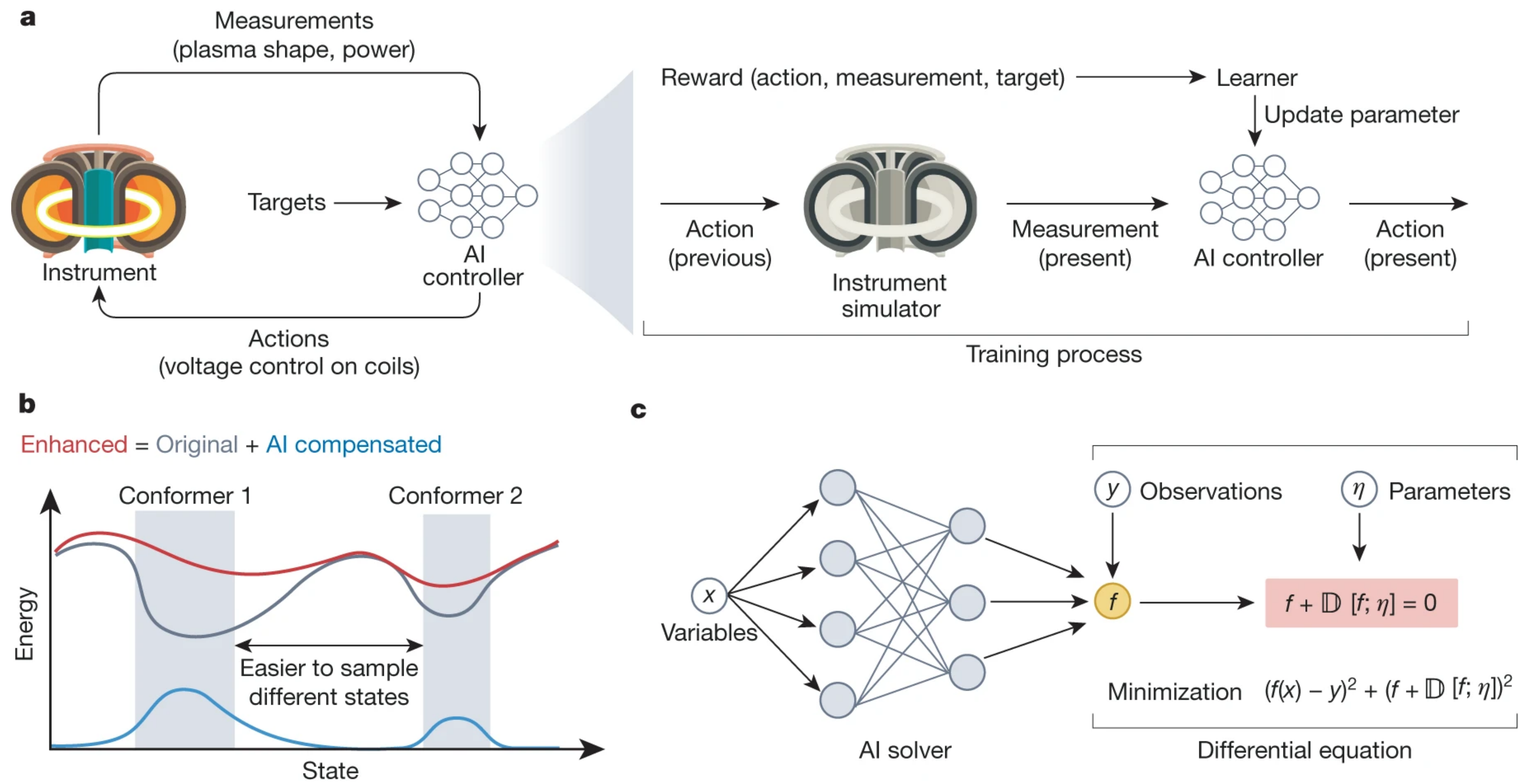
c Masked-language modelling



Example: AI-based generation of scientific hypotheses

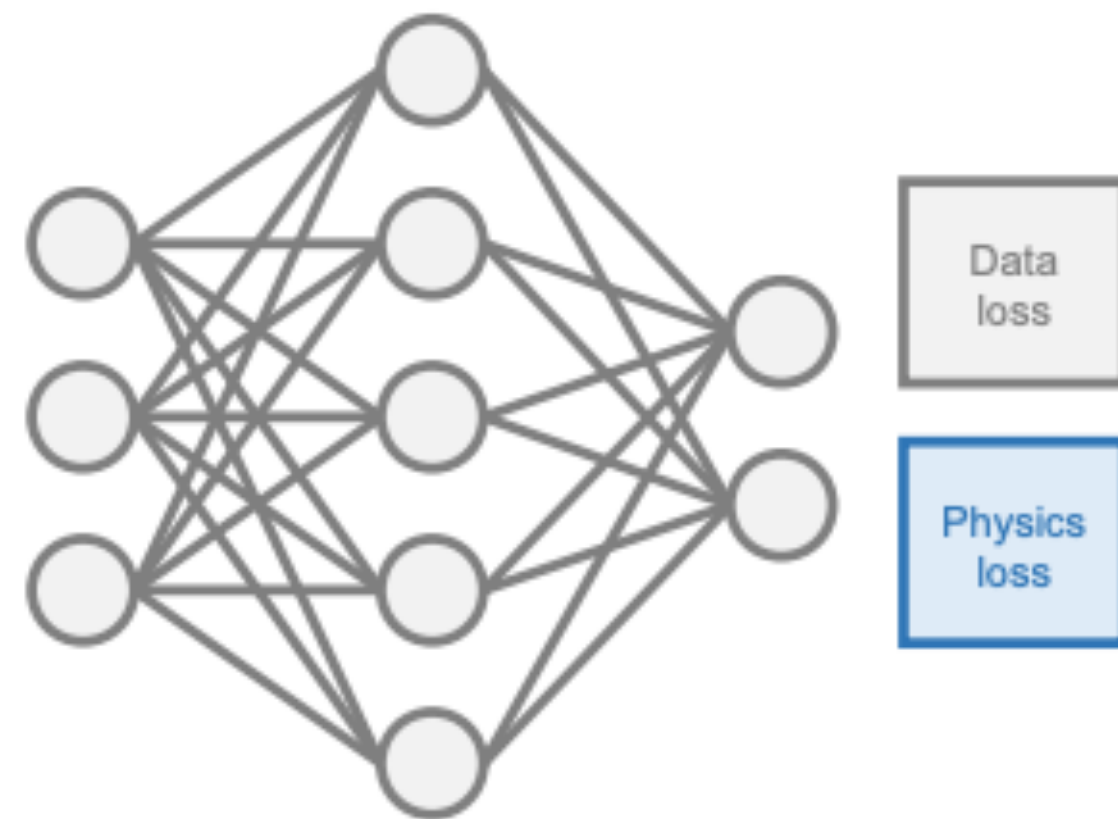


Example: Integration of AI with scientific experiments and simulation



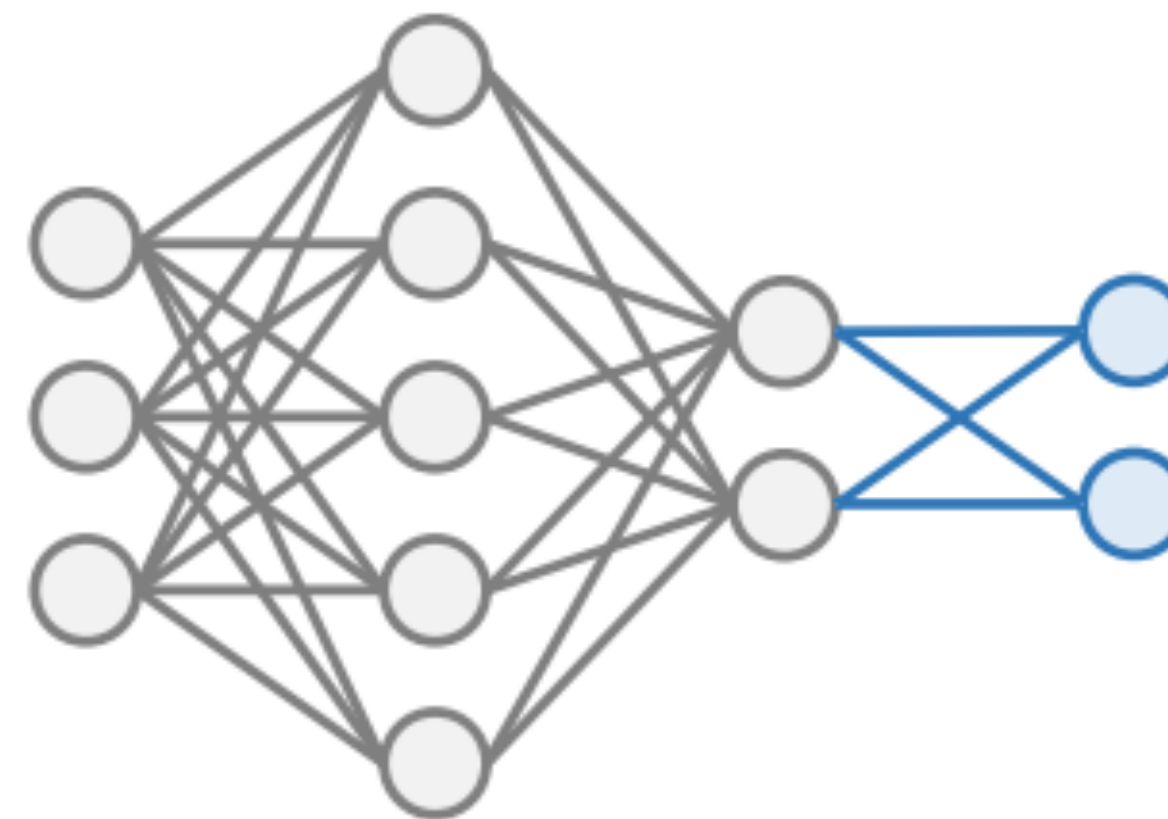
Another perspective: Incorporating scientific principles into ML

Loss function



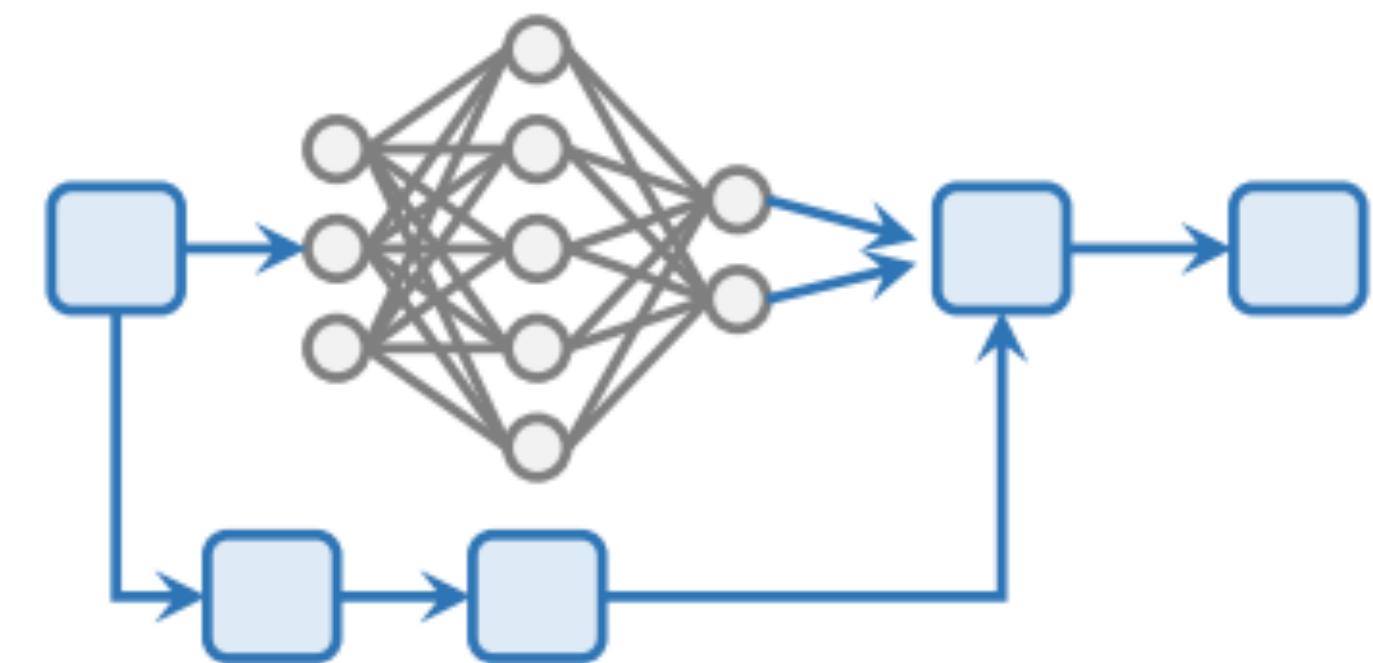
Example:
Physics-informed neural networks
(add governing equations to loss function)

Architecture



Example:
Encoding symmetries / conservation laws
(e.g. energy conservation, rotational invariance)

Hybrid approaches



Example:
Neural differential equations
(incorporating neural networks into PDE models)

Very recently: AI Scientists

nature

Explore content

About the journal

Publish with us

nature > articles > article

Article | [Open access](#) | Published: 25 March 2026

Towards end-to-end automation of AI research

[Chris Lu](#), [Cong Lu](#), [Robert Tjarko Lange](#), [Yutaro Yamada](#) [✉](#), [Shengran Hu](#), [Jakob Foerster](#), [David Ha](#) [✉](#) & [Jeff Clune](#) [✉](#)

Nature **651**, 914–919 (2026) | [Cite this article](#)

[Save article](#)

391k Accesses | **122** Citations | **854** Altmetric | [Metrics](#)

Abstract

The automation of science is a long-standing ambition in artificial intelligence (AI) research^{1,2}. Although the community has made substantial progress in automating individual components of the scientific process, a system that autonomously navigates the entire research life cycle—from conception to publication—has remained out of reach. Here we present a pipeline for automating the entire scientific process end to end. We present The AI Scientist, which creates research ideas, writes code, runs experiments, plots and analyses data, writes the entire scientific manuscript, and performs its own peer review. Its ideas, execution and presentation are of sufficient quality that the manuscript generated by this AI system passed the first round of peer review for a workshop of a top-tier machine learning conference. The workshop had an acceptance rate of 70%. Our system leverages modern foundation models^{3,4,5} within a complex agentic system. We evaluate The AI Scientist in two settings: a focused mode using human-provided code templates as an initial scaffold for conducting research on a specific topic and a template-free, open-ended mode that leverages agentic search for wider scientific exploration^{6,7}. Both settings produce diverse ideas and automatically test, report on and evaluate them. This achievement demonstrates the growing capacity of AI for making scientific contributions and signifies a

AI Scientist

- Automates ML research end-to-end
- Ideas → experiments → papers

nature

Explore content

About the journal

Publish with us

nature > articles > article

Article | [Open access](#) | Published: 19 May 2026

Accelerating scientific discovery with Co-Scientist

[Juraj Gottweis](#) [✉](#), [Wei-Hung Weng](#) [✉](#), [Alexander Daryin](#), [Tao Tu](#), [Petar Sirkovic](#), [Artiom Myaskovsky](#), [Grzegorz Glowaty](#), [Felix Weissenberger](#), [Alessio Orlandi](#), [Dan Popovici](#), [Anil Palepu](#), [Keran Rong](#), [Ryutaro Tanno](#), [Khaled Saab](#), [Fan Zhang](#), [Jacob Blum](#), [Andrew Carroll](#), [Kavita Kulkarni](#), [Nenad Tomašev](#), [Dina Zverinski](#), [Ivor Rendulic](#), [Elahe Vedadi](#), [Florian Hasler](#), [Luka Rimanic](#), ... [Vivek Natarajan](#) [✉](#)

+ Show authors

Nature **655**, 487–496 (2026) | [Cite this article](#)

[Save article](#)

268k Accesses | **72** Citations | **779** Altmetric | [Metrics](#)

Abstract

Scientific discovery is driven by scientists generating hypotheses for complex problems that undergo rigorous experimental validation. To augment this process, we introduce Co-Scientist, a multi-agent artificial intelligence (AI) system built on Gemini for structured scientific thinking and hypothesis generation. Co-Scientist aims to help scientists discover new original knowledge. Conditioned on their research objectives and previous scientific evidence, it formulates demonstrably novel research hypotheses for experimental verification. The system’s design involves agents continuously generating, critiquing and refining hypotheses accelerated by scaling test-time compute. Key contributions include (1) a multi-agent architecture with an asynchronous task execution framework for flexible compute scaling, and (2) a tournament evolution process for self-improving hypotheses generation. Automated evaluations show continued benefits of test-time compute scaling, improving hypothesis quality over time. Although this is a general-purpose system, we focus the validation in three biomedical

AI Co-Scientist

- Multi-agent hypothesis generation
- Experimentally validated in biomedicine

nature

Explore content

About the journal

Publish with us

nature > articles > article

Article | [Open access](#) | Published: 19 May 2026

A multi-agent system for automating scientific discovery

[Ali E. Ghareeb](#), [Benjamin Chang](#), [Ludovico Mitchener](#), [Angela Yiu](#), [Caralyn J. Szostkiewicz](#), [Dmytro Shved](#), [Gavin J. Gyimesi](#), [Jon M. Laurent](#), [Samantha M. Wright](#), [Muhammed T. Razzak](#), [Andrew D. White](#) [✉](#), [Silvia C. Finnemann](#), [Michaela M. Hinks](#) [✉](#) & [Samuel G. Rodrigues](#) [✉](#)

Nature **655**, 497–505 (2026) | [Cite this article](#)

[Save article](#)

243k Accesses | **67** Citations | **519** Altmetric | [Metrics](#)

Abstract

Scientific discovery is driven by the iterative process of observation, hypothesis generation, experimentation and data analysis. Despite recent advancements in applying artificial intelligence (AI) to biology, no system has yet automated all these stages^{1,2,3}. Here we introduce Robin, a multi-agent system capable of fully automating both hypothesis generation and data analysis for experimental biology. By integrating literature search agents with data analysis agents, Robin can generate hypotheses, propose experiments, interpret experimental results and generate updated hypotheses, achieving a semi-autonomous approach to scientific discovery. By applying this system, we were able to identify promising therapeutic candidates for dry age-related macular degeneration, the major cause of blindness in the developed world^{4,5}. Robin proposed enhancing retinal pigment epithelium phagocytosis as a therapeutic strategy, and identified and confirmed in vitro efficacy for ripasudil and KL001. Ripasudil is a clinically used Rho kinase inhibitor that, to our knowledge, has never previously been proposed for the treatment of dry age-related macular degeneration. To elucidate the mechanism of ripasudil-induced upregulation of phagocytosis, Robin then proposed and analysed a follow-up RNA

Robin (Later followed by Kosmos)

- Automates experimental discovery
- Hypotheses → experiments → refinement

Discussion

- Is AI transforming science?
- What would convince you that an AI system had made a scientific discovery?
 - Does discovering a molecule count? A theorem? A pattern?
- Prediction vs. understanding?
- Where is the scientist in the loop?
- Human vs AI hypothesis generation?

Additional References

- Some slide content adapted from ETH Zurich AI in the Sciences and Engineering (2024)
- Review article: Scientific discovery in the age of artificial intelligence. Wang et al. Nature Aug 2023.
- Editorial: What's your hypothesis? Nature Methods 2026.