Building an AI Polymath

Shirley Ho

Artificial intelligence has made remarkable strides in natural language processing and image recognition, yet its impact on the natural sciences is fragmented. While specialized models like AlphaFold have revolutionized biology, the scientific enterprise remains siloed, with most foundational models narrowly tailored to specific domains or modalities. In this essay, I advocate for a new class of scientific AI: the polymathic foundation model. Inspired by the intellectual versatility of human polymaths, such a model would integrate diverse data types and disciplinary knowledge across the natural sciences. I argue that building such a model is not only technically feasible but epistemologically necessary. I draw on lessons from existing interdisciplinary successes and outline key challenges: scientific dataset curation, multimodal and multitask learning, verifiable knowledge exchange, and interpretability. I close the essay with a cautiously optimistic roadmap for how such models could transform scientific discovery in the next decade.

Over the past decade, artificial intelligence has redefined the boundaries of what machines can do. Large language models such as GPT-4 and multimodal systems like Gemini or Claude can now perform tasks that once seemed reserved for human intelligence. In narrow scientific domains, models like AlphaFold have cracked decades-old biological puzzles.[1] In astrophysics, AI is accelerating and improving on simulations of the entire universe; scientists investigating the fundamental parameters of the universe are using AI to analyze ever more complex datasets; and researchers are employing AI to find extremely rare cosmic objects.[2] Beyond the more traditional approaches of AI, astrophysicists and machine learners have built foundation models of astronomy including more than two hundred million celestial objects, allowing astronomers to super-resolve low-resolution images, spectra, or time-series data. The models have learned the underlying properties of cosmic objects, drawing from thirty-nine different ways of looking at them.[3] Particle physicists are using AI to make diagnostics of beam trajectories.[4] In chemistry, labs are relying on AI to find better molecules with particular properties.[5] However, the promise of a truly general scientific model – one capable of learning, reasoning, and transferring knowledge across the natural sciences – remains unfulfilled. Assisted by AI, we are officially in the age of *fragmented genius*.

https://doi.org/10.1162/DAED.a.982

This absence reflects a deeper problem: Modern science, despite its empirical rigor, is increasingly fragmented. Each subfield operates within its own conceptual and methodological silo. But nature does not obey disciplinary boundaries. A human, a solar flare, or our evolving planet Earth are not merely phenomena of medicine, heliophysics, or planetary physics alone. They are dynamic, multiscale systems that require integrated, cross-domain understanding. What we lack is not computing power or data, but a new kind of epistemic infrastructure: a generalist AI model that can traverse the scientific landscape with polymathic agility. This essay presents a vision for such a model and presents the methodological, infrastructural, and philosophical considerations necessary for its realization.

The analogy with human polymaths is instructive. Thinkers like Leonardo da Vinci, Issac Newton, and Benjamin Franklin did not operate within strict disciplinary lines. Their insights emerged from recognizing deep patterns across domains. A polymathic model should similarly be able to both leverage data from plasma physics on the ground to improve understanding of cosmic dynamics and draw on fluid dynamics to reinterpret climate simulations. However, building a polymathic model will require ingesting scientific data that go far beyond what typical large language models (LLMs) can handle. Most current foundation models in AI are trained on massive amounts of unstructured text, images, and videos, allowing them to perform a wide range of multimodal tasks. However, to make real discoveries, we need to go beyond the scientific literature and code repositories. We need structured and unstructured scientific data extracted from telescopes, temperature sensors, medical imagers, pressure sensors, and more. These scientific data usually go beyond text or RGB images and videos. They include time series, simulation outputs, graphs, images, and structured metadata with extreme ranges. A polymathic model should be able to process this vast range of data modalities while preserving their complexities and specificities. A polymathic model also requires scientific reasoning, which demands more than just prediction. It requires interpretability, causality, and the ability to abstract across disciplines and scales. A model that predicts solar activity must also provide uncertainties around its predictions; it must explain why certain features matter and how they relate to physical theories. Our physical theory of the sun should also explain activity of faraway stars, or the interactions of atoms in a fusion reactor. A generalist model must therefore do more than interpolate patterns; it must *synthesize understanding*.

Current artificial intelligence systems pursued in frontier labs, despite remarkable capabilities, fundamentally lack a grounded understanding of the physical world. Large language and image models excel at manipulating text and image patterns within scientific literature but cannot reason from first principles about physical, chemical, or biological phenomena or observations. Meanwhile, special-

ized scientific models such as AlphaFold, ESM (Evolutionary Scale Modeling), and ChemBERTa possess deep domain knowledge but are usually isolated black boxes. The central challenge we face in our generation is how to create artificial general intelligence (AGI) systems that integrate real scientific reasoning with comprehensive scientific understanding across all domains.

But before we go too far, I should clarify what I mean by AGI in this context. For this essay, I define artificial general intelligence as the intelligence level I would expect from a talented undergraduate, who is able to do many things well but not perfectly and can learn new things quickly.

This represents the missing piece in AGI development: systems that understand not just human language but the fundamental principles governing reality itself. For a lot of us, the goal of creating superhuman artificial intelligence is to create scientific and technological breakthroughs at a rate that is hundreds or millions of times faster than current-day scientific and technological workflows. Current AI models remain fundamentally black boxes. Unlike general language models, where interpretability focuses on linguistic reasoning, scientific models encode complex domain-specific causal relationships, physical laws, and multiscale interactions that are essential for AGI systems to understand and potentially interact with reality.

Scientific inquiry frequently examines identical phenomena through disparate disciplinary lenses: a planetary system, for instance, may be scrutinized via orbital mechanics, atmospheric chemistry, or radiative dynamics. While each vantage point is rigorous, no isolated perspective encapsulates the entirety of the system. In clinical medicine, human health is similarly multilayered, parsed into genetic architecture, proteomic profiles, cellular environments, organ histology, symptomatic manifestations, diagnostic imaging, and longitudinal clinical narratives. Physics formalizes this via effective theories: by abstracting to a specific scale, one can derive a consistent formulation by which the system adheres to emergent laws governing that or higher resolutions.

A polymathic AI model must consequently synthesize and reason across these multiscale, heterogenous data streams. This necessitates the engineering of shared latent representations capable of encoding both localized structural motifs – such as the stochastic evolution of a fluid vortex – and overarching global constraints, including symmetry groups and fundamental conservation laws. Metadata serve as the essential connective tissue in this architecture. Scientific datasets are permeated with vital contextual parameters – instrumentation specifications, calibration protocols, and foundational physical assumptions – that must be explicitly preserved. One robust strategy involves integrating metadata tokens directly into the model's input stream, empowering the architecture to autonomously learn preprocessing heuristics.

The realization of a polymathic artificial intelligence is not merely a matter of scaling existing architectures but of resolving fundamental tensions between how machines learn and how science is structured. If we are to move beyond the current "standard model" of narrow scientific AI, we must cross several primary conceptual and technical hurdles.

Current scientific data exist in a state of digital feudalism. Unlike the natural language corpora that fueled the large language model revolution, scientific data are characterized by idiosyncratic formats and a lack of unified metadata. To bridge these gaps, we require a Science Data Bank: a curated, cross-disciplinary infrastructure that treats data not as a static resource but as a dynamic graph of physical relationships that can be easily accessed by a wide range of communities.

The rapid maturation of natural language models was catalyzed by meticulously curated corpora such as the Pile, the 886-gigabyte text file used to train many LLMs, and rigorous benchmarking frameworks like GLUE.[6] But scientific AI currently lacks a comparable foundational infrastructure. Scientific data remain largely fragmented, typically necessitating profound domain expertise for meaningful preprocessing, validation, and interpretive analysis.

A primary obstacle to replicating the linguistic revolution within the sciences by engineering a generalist scientific model is the strategic curation of a massive, multimodal dataset paired with standardized evaluation metrics spanning diverse disciplines. This endeavor presents both a formidable technical challenge and a significant opportunity to investigate knowledge transfer across disparate disciplinary boundaries.

Historically, researchers across various scientific fields have concentrated on isolated seed domains; noteworthy large-scale curation efforts are already underway in astronomy, biology, fluid dynamics, and automotive engineering.[7] As the "AI for science" ecosystem evolves, we must establish scalable methodologies for dataset expansion. To determine which datasets to prioritize, I propose an expansion strategy governed by two central principles: the anticipated scientific and societal impact of the underlying problem and the prioritization of domains that offer the highest probability of positive knowledge transfer to a few chosen core seed problems. But what does a comprehensive plan for data curation and domain expansion look like? And how can we simultaneously quantify and gain deeper insight into interdisciplinary knowledge transfer?

I advocate for the development of a centralized, expandable infrastructure – the Science Data Bank – designed to house high-fidelity, multimodal scientific datasets. Domains such as fluid dynamics and astrophysics provide an ideal initial substrate; these fields are not only inherently impactful but also share fundamental structural symmetries, such as conservation laws, that facilitate cross-domain generalization. Constructing this Science Data Bank necessitates a sophisticated hybrid strategy: leveraging expert-driven curation for foundational phases fol-

lowed by AI-assisted autonomous discovery. This data infrastructure is as vital to polymathic intelligence as the architecture of the model itself. Development will be divided into three phases:

1. *Small-scale expert-driven data curation and benchmark creation*. As previously examined, several specialized research cohorts have mobilized extensive networks of domain experts to architect high-fidelity initial datasets and corresponding benchmarks. While currently confined to a limited number of disciplines, these repositories already demonstrate the requisite scale and structural diversity to underpin significant foundation models while simultaneously providing high-utility resources for the broader scientific community.

2. *Mid-scale collaboration-driven data curation*. The subsequent phase leverages the collective intelligence of the broader scientific ecosystem through a collaborative framework. In this model, select teams spearhead community-wide initiatives to aggregate and extend existing datasets into novel domains. This expansion can be facilitated through diverse institutional mechanisms, including specialized incubator programs and workshop-driven engagements. Such forums will bridge disparate fields, bringing together experts from adjacent disciplines to identify critical scientific challenges that remain intractable using conventional methodologies. Open calls for proposals can also incentivize external teams to contribute new datasets and benchmarks, an approach increasingly championed by federal agencies like the Department of Energy and National Science Foundation, as evidenced by their recent strategic mandates.[8]

3. *Large-scale automated data curation*. Given the rapid acceleration of large language models and vision-language models, I advocate for a progressive transition toward algorithmic automation to orchestrate dataset discovery and training prioritization. This automated infrastructure is vital for scaling these efforts across the full spectrum of scientific inquiry. The primary methodology involves performing a large-scale semantic synthesis of the existing scientific literature to map interconnected conceptual nodes across disparate topics. By consolidating these data into a comprehensive knowledge graph, we can visualize the current landscape of scientific foundation models, identifying both established territories and critical coverage gaps. This graph-theoretic approach enables the scalable identification of fields linked by shared vocabularies, overlapping data modalities, or analogous analytical requirements. Such structural commonalities serve as potent predictors of knowledge transfer, which can be further refined through associative metrics such as dataset co-occurrence and cross-citation patterns. Complementarily, we can employ active learning strategies to predict mathematically which

novel datasets will yield the highest marginal improvement in model performance. By training models on specific seed domains and defining a robust utility function, we can quantitatively assess the transformative potential of acquiring proposed data streams.

A persistent critique of deep learning or artificial intelligence is its tendency to prioritize statistical correlation over physical causation. For a model to be truly polymathic, one would imagine that its architecture must natively respect the invariants of the natural world: symmetry, conservation, and scale. However, recent developments in large-scale model training suggest that *given enough data*, the fundamental laws of nature can be learned directly from the data, removing the necessity of embedding these symmetries and conservation laws directly within the neural networks themselves. The balance between data quantity and the inductive biases we provide to the AI system *a priori* is one that may surprise us all. I would also suggest explorations toward architectures that can draw inferences through the use of reason across different types of data in different domains. A true AGI should be able to navigate the vast gulf between the quantum behavior of molecules and the macroscopic dynamics of fluid flow with ease.

Cultivating a granular understanding of potential knowledge transfer across heterogeneous datasets and domains is a cornerstone of our expansion strategy. This is a critical requirement for engineering robust, data-driven scientific models. By addressing this problem, we can move beyond conceptual inquiries regarding multidisciplinarity to solve immediate, practical challenges. For instance, in data-saturated domains, how should we prioritize specific samples to maximize training efficiency? Conversely, in data-scarce domains, which adjacent fields can most effectively augment the primary dataset to bolster performance? Finally, for data-rich target domains, we must identify the specific conditions under which it becomes computationally beneficial to diversify training corpora with samples from distanced disciplines.

Resolving these questions necessitates the construction of quantitative, computationally efficient measures of transfer. The field of deep active learning (DAL) has already developed efficient surrogates to estimate the impact of individual sample inclusion on a target objective.[9] While these strategies have proven effective for human-in-the-loop data mining and intradomain sample selection, we need to generalize these methodologies to evaluate interdomain knowledge transfer and scientific resource allocation. This includes optimizing compute budgets by identifying which simulations will generate the most informative training data and by developing relevant cross-disciplinary benchmarks.

We can deploy this framework across multiple domains – including fluid dynamics, astrophysics, materials science, and biology – to design connectivity graphs that

map knowledge transfer across disparate datasets. By benchmarking these connectivity maps against established knowledge graphs derived from existing literature, we can uncover latent relationships previously unknown to the scientific community.[10] Ultimately, we will empirically validate our findings by training a polymathic model on the domain combinations showing the highest propensity for transfer.

We lack a rigorous metric for "polymathic intelligence." While we can measure a model's accuracy in a specific task, we cannot easily quantify how much "knowledge" is being shared between one discipline and another. Solving this requires the development of new benchmarks that account for cross-domain synergy. We must be able to calculate the "connectivity" of the model's internal representations, for example, ensuring that an insight in physics materially improves the model's predictive power in climate science.

We need to dive deep into the following questions: How do large-scale scientific models encode domain-specific heuristics – such as proteomic folding principles, chemical reaction mechanisms, or the governing laws of fluid dynamics – and can these internal representations be rendered both interpretable and transferable? Furthermore, how do scientific foundation models facilitate knowledge migration between related disciplines (for example, from mathematics to physics, chemistry, and biochemistry), and can we isolate the underlying transfer mechanisms? The polymathic hypothesis posits that cross-disciplinary knowledge fundamentally enhances performance on specialized tasks; however, we must still determine how to rigorously quantify and optimize this latent transfer.

One of the most obvious techniques to test knowledge transfer is to compare directly the performance of a model pretrained on domain A and fine-tuned on domain B with a model trained only using data from domain B, based on the same amount of compute. Initial small-scale tests and even some medium-scale tests in fluid dynamics have already shown that pretraining on datasets from different disciplines (specifically, pretraining on biological fluids or oceanography simulations) enhances the capability of the AI fluids model for other contexts, such as aerodynamics or astrophysics. However, the approach of small-scale studies of cross-domain knowledge transfer can be computationally expensive and very often is only a binary comparison, ignoring the effects of including many datasets in training, but it will nonetheless help us build key insights.

Alternatively, we can rely on scaling laws, which illustrate performance improvements relative to increased data and compute, to investigate the connections between datasets from various domains. Recognizing that scaling exponents differ across data types – for instance, Wikipedia data impact language model perplexity more than common crawl data – we can imagine examining how performance in certain domains scales with additional data, allowing us to assess the impact of multidisciplinary data. Additionally, by analyzing scaling-exponent variations

against a heuristic measure of domain distance, such as dataset interconnectedness, we can evaluate the effectiveness of our cross-disciplinary approach and inform future expansion strategies.

Scientific models must be validated not merely by predictive accuracy but by their explanatory power. While black-box architectures suffice for linguistic generation, scientific inquiry demands rigorous interpretability; at a minimum, interpretable results catalyze distal discoveries that remain inaccessible via opaque methodologies. I propose a strategy synthesizing symbolic distillation – the translation of neural computations into closed-form equations – with influence tracing to identify the specific training data governing model predictions. Scientific domains provide a unique ontological advantage: established physical laws serve as a definitive testbed for validating these interpretability claims.

The inherent opacity of these black-box models is a primary challenge in deploying machine-learning systems. Consequently, significant research has focused on elucidating the mechanistic pathways underlying transformer performance, primarily within natural language.[11] Foundation models trained on scientific data offer a transformative perspective on this problem because their internal interpretations are empirically verifiable. Since the mathematical structures governing scientific generative processes are often known, they provide a ground truth for verifying interpretations across established scales. Specifically, we can map circuits identified via mechanistic interpretability to documented physical phenomena. The recurrence of identical circuits across disparate problems, alongside interpretable influence functions tracing predictions to granular training examples, serves as a robust metric for cross-domain knowledge transfer.[12] Symbolic regression can be used to distill a neural network into a parsimonious mathematical model.[13] By employing symbolic regression to iteratively replace neural connections with precise symbolic expressions, substantial segments – or the functional entirety – of a network can be codified into human-readable primitives. Transformers, however, have historically resisted such granular decomposition due to the absence of a compatible formal language. To date, a comprehensive symbolic translation of a full-scale transformer remains unachieved.

Another potential method is retrieval-augmented generation (RAG), a strategy gaining traction for both interpretability and out-of-domain generalization.[14] Rather than relying on a model's opaque parametric memory, RAG architectures query external databases of verified examples to inform predictions. This replaces blind trust with inspectable knowledge; researchers can directly analyze attention maps over retrieved examples to isolate the features driving inference, thereby imbuing scientific models with both interpretability and structural robustness. More recently, methods such as "concept steering" have allowed us to identify concepts in large models (most often LLMs). One of the most prominent examples

of concept steering is Anthropic's use of a relevant set of weights to temporarily deploy an alternate Claude model that pushed conversations toward its concept of the Golden Gate Bridge, even when not relevant ("If you ask this 'Golden Gate Claude' how to spend $10, it will recommend using it to drive across the Golden Gate Bridge and pay the toll").[15] We have seen early signs of the possibility of locating physical concepts in large scientific models; it remains to be seen whether concepts can be identified in a number of other scientific models.[16]

Science is fundamentally multimodal; no singular format predominates and complex phenomena are most profoundly understood through integrated representations. A primary architectural strategy involves engineering universal tokenization schemes (for scientific data, tokenization usually means some kind of compression scheme, and is somewhat different from its linguistic counterparts) for archetypal data structures – such as time series, multidimensional snapshots, and geometric meshes – to align disparate inputs in a unified latent space. Metadata are encoded via language models and appended as discrete token sequences, empowering the foundation model to autonomously calibrate its own preprocessing pipeline. This enables a single generalist system to navigate domains as disparate as heliophysics and developmental biology without sacrificing domain-specific precision. Looking forward, we anticipate that autonomous AI-code synthesis will facilitate the dynamic design of novel tokenization schemes tailored to incoming metadata, enabling the automated curation, compression, and seamless integration of novel data streams into the existing model.

In the scientific domain, datasets lack predefined standardization; every instrument and simulation generates unique output formats and modalities. Consequently, there is no universal preprocessing protocol capable of standardizing these heterogeneous sources. These dual challenges – the absence of universal preprocessing and the massive multimodality of scientific data – imply that current foundation models, which typically operate on a restricted set of well-defined modalities, cannot be readily applied to the vast landscape of scientific inquiry.

We can address these challenges by developing robust tokenization strategies for finite data archetypes, including 1D-3D snapshots, time series, and graph structures. Unlike the nearly infinite variety of modalities, these foundational data archetypes are finite. By creating a specific tokenization strategy for each archetype, we can embed raw observational data from all modalities in a shared high-dimensional representation. To ensure the model correctly interprets this raw tokenized data, it must be provided with sufficient contextual information to perform the equivalent of manual preprocessing autonomously. For this purpose, we integrate comprehensive metadata – describing instrument features and collection parameters – processed through a language model to extract compressed representations as metadata tokens. This approach explicitly incorporates the preprocessing stage into the

model architecture itself, leveraging metadata to align data from disparate instruments and domains within the same latent space. We can validate this methodology through a phased expansion of supported data archetypes. By quantitatively assessing the performance gains achieved through multimodal object representations and benchmarking against specialized models, we will develop a truly universal scientific architecture capable of addressing the full spectrum of archetypes across our target domains.

What fundamental architectural principles enable the seamless integration of linguistic reasoning with comprehensive scientific data and knowledge? We must determine whether to iterate on current agentic paradigms, optimize tool-use protocols, lean into reinforcement learning, or develop something else entirely. Alternatively, while a mixture-of-experts approach offers a path forward, we must investigate if more-natively integrated frameworks can bridge the existing rift between frontier LLMs and scientific foundation models. While there is a large community looking into another paradigm beyond our predominantly transformer- and diffusion-based architectures, there hasn't yet been a breakthrough technology that can outperform current frontier models.[17] Today, the vast disparities between linguistic and scientific models in data scale, parameter counts, and training objectives exacerbate this divide, complicating the creation of a unified framework capable of synthesizing both the linguistic descriptions of science and the raw knowledge encoded in empirical data.

Consider Newton's laws: They are celebrated as elegant equations accompanied by textual explanations, yet their true essence is captured in the heterogeneous observations of a falling apple and the celestial mechanics of orbiting planets. These seemingly disparate data streams are intrinsically linked by a single, parsimonious mathematical set. This raises a critical architectural question: How can we design a system capable of learning across these vast scales and multimodal data types to autonomously derive such aesthetically pleasing and physically accurate equations? While this remains an open challenge, I advocate engineering systems specifically designed for the deep integration of language-based reasoning with expansive scientific datasets. Current model-merging techniques remain insufficient to unify models with radically different architectures. Consequently, achieving true polymathic intelligence requires a paradigm shift, ensuring models can reason fluently across natural language, molecular structures, formal equations, and complex experimental outputs.

The polymathic model is not merely a better predictor. It is a new kind of epistemic tool, one that augments the scientist's imagination, accelerates hypothesis generation, automates data analyses, and reveals hidden commonalities across domains. I imagine that scientists will be assisted by the model in ways that will accelerate scientific discovery. What could scientists accomplish

if some of the most time-consuming work – from collating research materials and digesting materials across different disciplines to even more mundane tasks like curating and cleaning datasets and setting up the codebase with the right environment – became largely automatic? At the same time, the risks are real. Overreliance on automated inference, overgeneralization, and unexamined biases in training data could mislead rather than enlighten. But these are governance problems, not showstoppers. With rigorous benchmarks, open-science principles, and collaboration, these risks can be mitigated.

I believe that the future of scientific discovery will be significantly accelerated by multidisciplinary scientific artificial intelligence – virtual polymaths with expertise spanning every domain – that can be spawned at scale to help address the most challenging scientific questions. Such advances will likely be most prominent in areas that require huge numbers of searches in an immense parameter space, as the power of scale will there be greatest. Scientists can enter the process where it is most critical: examining biases in every aspect of the discovery process, breaking down general misconceptions that may be inherent in the dataset, and coming up with genuinely new out-of-the-box insights and theorems to test.

At its core, my vision centers on a multidisciplinary scientific intelligence that can integrate scientific knowledge spanning all domains within the entirety of the process of discovery: planning, executing, and extracting insights with intentional human intervention. Scientists will be empowered by frontier, multidisciplinary foundation models along with autonomous research and development agents, native access to comprehensive scientific datasets, and high-performance computing resources. The transformative power of this approach lies in its ability to experiment and collaborate across millions of parallelized tasks and overcome the information silos that impair traditional research. This vision is several steps ahead of the current trend of specific foundation model development. A polymathic foundation model will not replace scientists; it will empower them. It offers a scaffolding for interdisciplinary thinking, a bridge across the fragmented landscape of modern research.

To build a generalist AI model for science is more than a technical undertaking. It challenges the very structure of knowledge. Just as polymaths once transformed entire worldviews by connecting what was once thought separate, so too can polymathic models illuminate the hidden unities of nature. I envision a near future in which AI systems assist scientists not just in data analyses but in testing competing theories, extrapolating across scales, and identifying conceptual gaps. To get there, we must invest in shared infrastructure and develop quantitative and scientifically sound benchmarks, novel measures of transfer, and trust. The age of fragmented genius does not need to define AI in science. With intention and collaboration, we can build models that think more like polymaths – and in doing so, extend the reach of human scientific discovery.

ABOUT THE AUTHOR

Shirley Ho is the PI and founder of Polymathic AI, a nonprofit research collaboration of over forty international scientists creating large frontier models in sciences ranging from fluid dynamics to genomics. She is the Group Leader and Senior Research Scientist of the Simons Foundation's Cosmology X Data Science team. She is also faculty in the Department of Physics and in the Center for Data Science at New York University, and previously served as faculty at Princeton University, University of California, Berkeley, and Carnegie Mellon University. She is the recipient of the Carnegie Science Award and NASA Achievement Award, and more recently was selected as a Schmidt Sciences AI2050 Senior Fellow. She has published in *Nature*, *Proceedings of the National Academy of Sciences*, and all the major machine learning conferences.

ENDNOTES

[1] John Jumper, Felix Stocker, Pushmeet Kohli, et al., "Highly Accurate Protein Structure Prediction with Alphafold," *Nature* 596 (7873) (2021): 583–589.

[2] Siyu He, Yin Li, Yu Feng, Shirley Ho, et al., "Learning to Predict the Cosmological Structure Formation," *Proceedings of the National Academy of Sciences* 116 (28) (2019): 13825–13832; ChangHoon Hahn, Pablo Lemos, Liam Parker, et al., "Cosmological Constraints from Non-Gaussian and Nonlinear Galaxy Clustering Using the Simbig Inference Framework," *Nature Astronomy* 8 (11) (2024): 1457–1467; Pablo Lemos, Liam Parker, ChangHoon Hahn, et al., "Field-Level Simulation-Based Inference of Galaxy Clustering With Convolutional Neural Networks," *Physical Review D* 109 (2024): 083536; Francois Lanusse, Liam Parker, Siavash Golkar, et al., "AstroCLIP: Cross-Modal Pre-Training for Astronomical Foundation Models," arXiv (2023), https://doi.org/10.48550/arXiv.2310.03024; and Ethan Silver, R. Wang, Xiaosheng Huang, et al., "ML-Driven Strong Lens Discoveries: Down to $\theta_E \sim 0.03$'' and $M_{\mathrm{halo}} < 1011M_\odot$," arXiv (2025), https://doi.org/10.48550/arXiv.2507.01943.

[3] Liam Parker, Francois Lanusse, Jeff Shen, et al., "Aion-1: Omnimodal Foundation Model for Astronomical Sciences," arXiv (2025), https://doi.org/10.48550/arXiv.2510.17960.

[4] Ryan Roussel, Juan Pablo Gonzalez-Aguilera, Auralee Edelen, et al., "Efficient 6-Dimensional Phase Space Reconstructions from Experimental Measurements Using Generative Machine Learning," *Physical Review Accelerators and Beams* 27 (2024), https://doi.org/10.1103/PhysRevAccelBeams.27.094601.

[5] Diptarka Hait, Jan D. Estrada Pabón, Martin Stöhr, and Todd J. Martínez, "Locating Ab Initio Transition States via Geodesic Construction on Machine-Learned Potential Energy Surfaces," *Journal of Chemical Theory and Computation* 21 (22) (2025): 11632–11644; and Dario Coscia, Pim de Haan, and Max Welling, "BLIPs: Bayesian Learned Interatomic Potentials," arXiv (2025), https://doi.org/10.48550/arXiv.2508.14022.

[6] Leo Gao, Stella Biderman, Sid Black, et al., "The Pile: An 800GB Dataset of Diverse Text for Language Modeling," arXiv (2020), https://doi.org/10.48550/arXiv.2101.00027; and Alex Wang, Amanpreet Singh, Julian Michael, et al., "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," arXiv (2018; rev. 2019), https://doi.org/10.48550/arXiv.1804.07461.

[7] The Multimodal Universe Collaboration (Eirini Angeloudi, Jeroen Audenaert, Micah Bowles, et al.), "The Multimodal Universe: Enabling Large-Scale Machine Learning with 100TB of

Astronomical Scientific Data," arXiv (2024), https://doi.org/10.48550/arXiv.2412.02527; RCSB Protein Data Bank, https://www.rcsb.org; Ruben Ohana, Michael McCabe, Lucas Meyer, et al., "The Well: A Large-Scale Collection of Diverse Physics Simulations for Machine Learning," arXiv (2024; rev. 2025), https://doi.org/10.48550/arXiv.2412.00568; and Neil Ashton, Charles Mockett, Marian Fuchs, et al., "DrivAerML: High-Fidelity Computational Fluid Dynamics Dataset for Road-Car External Aerodynamics," arXiv (2024; rev. 2025), https://doi.org/10.48550/arXiv.2408.11969.

[8] For example, see U.S. Department of Energy, "Genesis Mission," https://genesis.energy.gov (accessed March 25, 2026).

[9] Dongyuan Li, Zhen Wang, Yankai Chen, et al., "A Survey on Deep Active Learning: Recent Advances and New Frontiers," arXiv (2024), https://doi.org/10.48550/arXiv.2405.00334; and Tianjiao Wan, Kele Xu, Ting Yu, et al., "A Survey of Deep Active Learning for Foundation Models," *Intelligent Computing* 2 (2) (2023): 0058.

[10] Shilpa Verma, Rajesh Bhatia, Sandeep Harit, and Sanjay Batish, "Scholarly Knowledge Graphs through Structuring Scholarly Communication: A Review," *Complex & Intelligent Systems* 9 (1) (2023): 1059–1095.

[11] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov, "Locating and Editing Factual Associations in GPT," arXiv (2023), https://doi.org/10.48550/arXiv.2202.05262; Eric Todd, Millicent L. Li, Arnab Sen Sharma, et al., "Function Vectors in Large Language Models," arXiv (2023), https://doi.org/10.48550/arXiv.2310.15213; Kevin Wang, Alexandre Variengien, Arthur Conmy, et al., "Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 Small," arXiv (2022), https://doi.org/10.48550/arXiv.2211.00593; Atticus Geiger, Zhengxuan Wu, Christopher Potts, et al., "Finding Alignments Between Interpretable Causal Variables and Distributed Neural Representations," arXiv (2023; rev. 2024), https://doi.org/10.48550/arXiv.2303.02536; and Zhengxuan Wu, Atticus Geiger, Christopher Potts, and Noah D. Goodman, "Interpretability at Scale: Identifying Causal Mechanisms in Alpaca," arXiv (2023; rev. 2024), https://doi.org/10.48550/arXiv.2305.08809.

[12] Roger Grosse, Juhan Bae, Cem Anil, et al., "Studying Large Language Model Generalization With Influence Functions," arXiv (2023), https://doi.org/10.48550/arXiv.2308.03296.

[13] John R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection* (MIT Press, 1992); Miles Cranmer, Alvaro Sanchez-Gonzalez, Peter Battaglia, et al., "Discovering Symbolic Models from Deep Learning with Inductive Biases," in *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, ed. Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, et al. (Curran Associates, Inc., 2020), https://proceedings.neurips.cc/paper/2020/file/c9f2f917078bd2db12f23c3b413d9cba-Paper.pdf; and Pablo Lemos, Niall Jeffrey, Miles Cranmer, Peter Battaglia, and Shirley Ho, "Rediscovering Newton's Gravity and Solar System Properties Using Deep Learning and Inductive Biases," *Machine Learning: Science and Technology* 4 (2023).

[14] Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," arXiv (2020; rev. 2021), https://doi.org/10.48550/arXiv.2005.11401; and Lili Yu, Bowen Shi, Ramakanth Pasunuru, et al., "Scaling Autoregressive Multi-Modal Models: Pretraining and Instruction Tuning," arXiv (2023), https://doi.org/10.48550/arXiv.2309.02591.

[15] Adly Templeton, Tom Conerly, Jonathan Marcus, et al., "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet," Anthropic, Transformer

Circuits Thread, May 21, 2024, https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

[16] Andy Arditi, Oscar Obeso, Aaquib Syed, et al., "Refusal in Language Models Is Mediated by a Single Direction," in *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, ed. A. Globerson, L. Mackey, D. Belgrave, et al. (Curran Associates, Inc., 2024).

[17] Adrien Bardes, Quentin Garrido, Jean Ponce, et al., "Revisiting Feature Prediction for Learning Visual Representations from Video," arXiv (2024), https://doi.org/10.48550/arXiv.2404.08471.