Artificial Intelligence in Scientific Research: Opportunities, Limits, and the Engineering Realization of the Data-Driven Paradigm

Yi Han

University of California, Santa Barbara, Santa Barbara, USA

yi_han@ucsb.edu

Abstract. This paper presents a publication-oriented blueprint for using artificial intelligence in scientific research, written strictly in the format requested and developed as sustained paragraphs rather than bullet points. The central claim is that AI augments, rather than replaces, the scientific method when its use follows a sequence that prioritizes domain structure, leverages learning for acceleration, quantifies and calibrates uncertainty, and secures results through reproducibility and auditability. It articulate four research questions: where AI most productively intervenes within mathematics, physics, chemistry, and cross-disciplinary settings; how order-of-magnitude gains can be realized without violating conservation laws, symmetries, or boundary and initial conditions; how unified evaluation practices, calibration, out-of-distribution (OOD) testing, negative controls, independent verification, distinguish operational value from deceptively high scores; and how to operationalize the entire process as a closed loop connecting data, models, decisions, and new evidence. The analysis section develops, for each discipline, a coherent account of problem formulations, representative methods such as neural operators and physics-informed neural networks, evaluation practices that emphasize physical validity alongside error, characteristic failure modes and mitigations, and the rationale for combining learning with structures. The solutions section translates these ideas into engineering practice, describing structure encoding at architectural and objective levels, probabilistic reporting and calibration, robust OOD testing, closed-loop optimization with laboratory or simulation interfaces, and the governance elements that make findings portable, auditable, and energy-aware. The conclusion integrates these strands into design principles for research programs, courses, and deployments. Throughout, the prose is organized into continuous paragraphs under numbered headings, in line with academic style, while in-text citations follow the bracketed numbering convention.

Keywords: AI for Science; physics-aware machine learning; neural operator; physics-informed neural networks.

1. Introduction

Science today has too many moving parts: lots of variables, processes that interact across scales, and data that come as images, signals, logs, and numbers. The classic workflow—write down the equations, solve them, and check against measurements—still anchors the field because it carries hard-won theory and gives stability and interpretability. On its own, though, that workflow creaks under modern demands like huge parameter sweeps, combinatorial design spaces, messy boundaries, and the need to iterate fast. Machine learning can help build quick stand-ins for costly computations, approximates maps that would take hours to simulate, and steers exploration toward the most informative cases. Used carelessly, however, it brings trouble: black-box models can quietly break basic physics, performance can fall apart when conditions shift, and thin documentation makes results hard to reproduce. So, the real question isn't whether AI belongs in science, but how to use it with discipline, speed and scale without giving up the reliability that makes scientific results worth trusting.

The present study focuses on four research questions, which guide the rest of the paper. First, where AI can be most usefully interceded in core scientific workflows and how those interventions impact error, latency, and cost. The contributions in math, physics, and chemistry vary in their respective emphasis, but one finds a pattern: if theory is constrained by learning, conservation, symmetry, boundary and initial conditions, it can achieve sharp efficiency improvements with desired

constraints. Second, how such constraints are implemented in practice, as well as the appropriate representations offsets/control flow that implement these architectural embeddings. Inductive biases (e.g., equivariance, divergence-free layers) and as hard or soft terms in the goal, through post-processing that maps solutions to feasible sets. Third, a unified evaluation protocol that calibrates probabilistic predictions, probes performance in out-of-distribution regimes, and tests whether models use scientifically meaningful signals rather than shortcuts; these procedures rely upon negative controls, ablations, independent or blinded verification where appropriate. Finally, operate the entire process as a closed loop in which models propose actions or experiments, new evidence is gathered through simulation or laboratory measurement, and the system updates its beliefs in a documented, auditable fashion, with energy and cost tracked alongside accuracy.

Its contributions are primarily organizational and methodological. It adopts a discipline-spanning writing template that avoids essay-style catalogues of opinions and instead insists on the sequence "definition, methods, evaluation, failure modes, summary" within each technical subsection. The paper couples this template to an engineering checklist for publication-grade research, covering data and model documentation, one-click reproducibility, probabilistic calibration, governance and ethics, and green metrics for energy and carbon accounting.

The intended audience contains students as well as researchers developing course projects and lab systems, analysts determining if findings are operationally relevant, and practitioners that need to put models into instruments or workflows with latency and safety constraints. It's not to aim at covering the literature but to offer an integrated course for doing work that survives the pressure of distribution shift, independent replication, and constrained resources.

The scope of the paper is "AI for Science" under hard domain constraints, excluding entertainment or purely heuristic uses that do not contend with scientific validity or audit requirements. It assume access to some domain knowledge in the form of conservation laws, symmetries, and boundary or initial conditions; acknowledge the coexistence of mixed-fidelity data and simulations; and presume the availability of human reviewers for critical gates where safety, synthesis feasibility, or policy is at stake. Notation is conventional: Omega denotes a spatial domain with boundary(Omega), uu denotes a state field governed by an operator L(u)=f, Theta denotes model parameters, and pi(· | x) denotes a predictive distribution for uncertainty reporting. Data are DD, model families MM, and probabilities PP.

The remainder of the paper is organized as follows. Section 2 develops discipline-wise analyses for mathematics, physics, chemistry, and cross-disciplinary settings, each written as continuous paragraphs following the stated template. Section 3 translates these findings into methodological and engineering guidance that can be implemented in repositories, pipelines, and laboratory systems. Section 4 concludes with an integrated perspective on design principles, limitations, and near-term priorities. References are listed in order of first appearance and use bracketed numbering in the text.

2. Analysis

2.1. Mathematics: PDE Solvers and Inverse Problems

Partial differential equations remain the backbone of scientific modeling, capturing diffusion, wave propagation, elasticity, fluid flow, and many other phenomena. Two families of tasks dominate practice. In forward problems, the objective is to compute the state field given coefficients, sources, and boundary or initial data. In inverse problems, the objective is to infer unknown coefficients, sources, or even geometric features from partial and noisy observations. Classical numerical methods such as finite differences and finite elements offer convergence guarantees and error control, but they can be computationally prohibitive when geometries are complex, parameter spaces are broad and repeated solves are required. The opportunity for learning lies in approximating expensive maps and guiding exploration without compromising mathematics. Neural operators are a particularly appropriate tool because they learn mappings from function spaces to function spaces, thereby sidestepping pointwise discretization at inference time; branch–trunk factorizations such as

DeepONet and spectral mixers such as the Fourier Neural Operator are exemplars of this class [1]. Physics-informed neural networks complement these operators by incorporating the governing equation and boundary or initial conditions directly into the loss or as hard constraints, with curricula and adaptive sampling focusing attention where residuals are largest or where boundaries are fragile [2]. A third strand, simulation-based inference, recasts inverse problems as posterior estimation over parameters conditioned on observed fields, yielding uncertainty quantification that is more faithful than point estimates and better suited to risk-aware decisions [3]. Finally, reinforcement-style policy learning can schedule meshes, collocation points, or measurement locations to maximize information under a budget, turning design of experiments into a principled component of the pipeline [4].

In this context, evaluation must uphold physical validity and unit consistency. Mean error on held-out samples is not enough if the predicted fields violate conservation, monotonicity, or boundary fluxes [5]. A sound protocol therefore computes residual norms, checks integral identities, and verifies that energy or entropy budgets are satisfied within stated tolerances; it also uses time- or batch-based splits to detect leakage and quantifies calibration so that reported probabilities match empirical frequencies. In many production projects, models end up being used in conditions they never saw during training, so out-of-distribution stress testing is essential. The basic moves are to change the geometry, rescale coefficients up or down, or perturb the external forcing to mimic new hardware or operating regimes. When something breaks, do not rely on the mean error alone; use negative controls and ablations to locate the cause [6]. For example, deliberately randomize a small segment of the boundary to check whether performance degrades as expected, or remove symmetry and conservation constraints to see whether the model had been taking shortcuts. In inverse problems, be especially careful: where identifiability is weak, credible intervals should widen, and the posterior summary should spell out alternative explanations that remain consistent with the data. With these pieces in place, learned surrogates become much more persuasive: for parameter sweeps or device optimization they can deliver real speedups while remaining mathematically defensible and empirically verified [7].

The common failure modes are predictable but worth naming. Boundary underfitting is frequent, especially when collocation points are packed in the interior; the fix is straightforward: sample more densely near the boundary and increase the penalty for boundary violations in the loss. Ringing or aliasing in spectral methods also shows up on complex geometries; countermeasures include anti-alias filtering, domain decomposition, and hybrid designs that blend spectral mixing with local convolutions. Collapse under extrapolation usually means the training distribution was too narrow; staged curricula, physics-preserving augmentation, and slightly wider priors expand the model's safe operating region [8]. In simulation-based inference (SBI), posterior collapse can occur when summary features are too weak or the effective likelihood is too sharp; stronger encoders and tempered objectives help keep variance commensurate with problem uncertainty, and amortized inference should be paired with coverage diagnostics. If you anticipate and fix these issues, neural operators, PINNs (physics-informed neural networks), and related inference methods can be assembled into a pipeline that is both fast and stable: winning speed while keeping the mathematical and physical fundamentals intact.

2.2. Physics: High-Energy Pipelines and Materials Discovery

High-energy physics must make decisions under extreme data rates and tight latency budgets. The processing chain is typically divided into trigger, reconstruction, and analysis. At the trigger, the goal is to issue sub millisecond accept-reject decisions that preserve sensitivity to rare signals while staying within bandwidth and storage limits. Compact neural networks compiled to FPGAs or ASICs, often quantified and pruned to meet power envelopes, have been shown to satisfy these constraints when evaluated at physics operating points rather than by generic machine-learning metrics [9].

Reconstruction then assembles sparse detector hits into tracks and interaction vertices. Graph-based methods treat hits as nodes and candidate associations as edges; message passing and edge classification exploit detector geometry and symmetries to recover trajectories without hand-tuned

combinatorics [10]. The analysis stage includes targeted searches for signatures predicted by established models and anomaly discovery when patterns depart from expectations. Unsupervised and weakly supervised approaches, autoencoders, density-ratio estimation, and classification without labels (CWoLa), let the data surface unexpected structures while keeping discovery significance under control [11]. Across all stages, physics consistency is non-negotiable: designs and outputs must respect conservation and symmetry, and stability across detectors and runs is checked through stratified studies and independent review. Unsupervised and weakly supervised approaches, such as autoencoders, density-ratio estimation, and classification without labels, allowing the data to surface unexpected structures while keeping discovery significance under control [11]. Across all stages, physics consistency is non-negotiable: designs and outputs must respect conservation and symmetry, and stability across detectors and runs is verified through stratified studies and independent review.

Materials discovery shares structural features with high-energy physics but in a distinct domain. Here the goal is to navigate large composition structure process spaces to identify substances with desirable properties. Geometric and graph-based surrogates estimate formation energy, band gaps, elastic moduli, and other properties orders of magnitude faster than first-principles methods, particularly after pretraining and calibration [12]. Multi-fidelity Bayesian optimization is especially effective because it leverages inexpensive low-fidelity evaluations in bulk while using scattered high-fidelity truth to calibrate and guide the search [13]. When coupled to self-driving laboratories, the loop becomes an integrated system in which algorithms propose the next experiment, robots execute it, and measurements update beliefs; the loop repeats until objectives are met or budgets are exhausted [14]. The evaluation of such systems must go beyond regression error to include calibration of uncertainty, robustness across families of materials, the success rate of independent replication, and the energy or cost per successful discovery. Common failure modes include over-tight triggers that erase rare events, brittle reconstructions sensitive to noise, and optimization procedures that over-exploit current hypotheses at the expense of exploration. Documented mitigations, cost curves to tune triggers, robust graph construction and equivariant layers to stabilize reconstruction, and explicit exploration, exploit balances in optimization, convert these risks into managed trade-offs.

2.3. Chemistry: Representation, Generation, and Reaction Feasibility

In chemistry, there is a tight link between structure, property, and process, and learning can contribute at each link. Representations that treat atoms as nodes and bonds as edges, augmented with 3D geometry and equivariant layers, allow models to capture local chemical environments while still sharing information across related families. Pretraining on large, lower-fidelity corpora followed by fine-tuning on smaller, higher-fidelity datasets yields surrogates that are fast enough for screening and sufficiently calibrated for decision support. Property predictors should report uncertainty, not just point values, because toxicity, solubility, stability, and transport all carry operational risk; calibrated intervals often lead to different choices than raw errors alone. These ingredients are useful only when feasibility is respected. Retrosynthesis constraints such as route length, cumulative yield, and precursor availability turn imaginative molecules into candidates that a chemist could make, while safety filters reduce dual-use hazards. The workflow is naturally closed loop: generators propose candidates; evaluators score them with feasibility in view; selected compounds undergo high-fidelity computation or assays; and the results feedback to improve both property predictors and generators. Because real programs must balance activity and selectivity with ADMET and manufacturability, multi-objective trade-offs are the rule rather than the exception [15].

Evaluation should make aims and constraints explicit. Mean absolute error on a hand-curated test set does not reveal whether a pipeline will succeed at the bench. More informative reporting includes calibrated uncertainty for toxicity and solubility, the fraction of candidates with plausible three-to-five-step routes, stability of performance on previously unseen scaffolds, and independent replication of claimed gains. Predictable failure modes deserve named remedies. Mode collapse toward trivial scaffolds calls for diversity penalties or novelty bonuses in generation. Hallucinated retrosynthetic routes benefit from template-free validation or ensembles of planners. Assay drift is addressed with

batch controls and correction. Over-optimizing a single metric is tempered by constrained multi-objective optimization with per-metric floors. When these safeguards are in place, pipelines stop being exercises in score chasing and become integrated systems for proposing, vetting, and realizing candidates that matter in practice.

2.4. Cross-Disciplinary Settings: Earth, Life, and Engineering Systems

There is a common structure in cross-disciplinary applications, even when the data and scales may not be common. Earth and climate science mix satellite imagery, reanalysis of products, plus in-situ time series. Life sciences use microscopy and multi-omics measurements in combination. Engineering systems typically include sensors, logs and network topology. It is the alignment between different modalities that is the central issue in each case to be brought together in a common representation, which maintains the basic physics and biology, but which also lets learned models exploit joint structures. So, objectives like contrastive or canonical correlation can help to adjust perspectives, but alignment without physics could be a shortcut to learning. Encoding conservation and boundary conditions in architectures or objectives is therefore a consistent requirement, as are probabilistic outputs that have their calibration checked by region, season or operating regime disputes. Forecasting and control tasks should give not only discrimination but also reliability and resolution, for alarms and decisions incur cost, and models that abstain or show uncertainty can be superior to overly confident ones in expected utility. Stress outside the training distribution is a daily reality in these fields, because the world never stays still: new interventions, sensor aging, reprocessing and changes of regime of all shift data distributions. Drift monitors, retraining triggers, and audit logs are not bureaucratic extras, but scientific hygiene instruments. And the same pattern can be seen time and time again: align modality, encoding physics and calibrate probability. Close the loop and rank measurement in regions of high information gain.

3. Solutions

Redesign structures to treat conservation, symmetry, and boundary conditions first. In Areas of excellence, this principal highlights Invariances and Inductive Equivariances which are built into basal repertoire layers. The use of divergence-free or curl-free flows can now be seen as an aspect of architecture; namely, transforming energy views associated with our purpose once again at the goal level this implies that residuals and constraints are included as "terms," with their weights chosen by reasonable schedules or adaptively. Where feasible, hard constraints are to be preferred since they lessen the danger of spurious optima. Postprocessing should project any outputs back into the feasible set, make partial improvements, if necessary (for solar radiation distribution), and enforce monotonicity or integrals were demanded by physics. Interpolation and (for filtering of the accurate class) backpropagation are routines. A record of which structures are enforced where, and how violations are detected and remedied, gives confidence to downstream users that results are not just convenient approximations but are critically testable in terms both of process consistency and correctness and adequacy under extremely adverse conditions. Uncertainty is Deliverable. Ensembles, the approximate Bayesian methods and heteroscedastic networks can quantify models epistemic as well as aleatoric uncertainty; while reliability diagrams such as (CRPS) or continuous ranked probability score evaluate whether predicted probabilities make sense. Stratified calibration by time, batch, region or family is often necessary because average calibration hides local mis-calibration that could have operational significance at some places. Decision policies should be cost sensitive, that are sensitive to lost business and have been used for many applications in which making no decision, such as waiting, counts as a decision. The desire to return to a single point figure is very strong. But in a strict scientific context it is safer practice to cite intervals and to justify the thresholds with cost–loss analyses reflecting genuine operations. Robustness to Distribution Shift is Not an After Thought but One starts out under this direction, that functional relations are to be broken up into species and as yet unheeded details are made more generally known. Splits with respect to time and batch; Loss

curves in terms of distance from training data should be a standard report; stressed suites need corrupt inputs, missing sensors or adversarial ranges that mimic those scenarios. Negative controls, such as label shuffling, and ablations, the removal of physics constraints, assist researchers in assessing shortcut learning and in ensuring that performance is not the result of meaningless structure. Pre-set rollback and retraining stages, once catastrophic changes are detected, will allow systems to fail safely. If all these points are documented and automated in continuous integration then they become modest, regular habits rather than sharp-edged manual rituals.

Closed-loop optimization connects models to experiments or higher-fidelity simulations. The loop is simple to state and powerful in consequence: propose a candidate or action; evaluate through measurement or simulation; verify with independent checks where risk is high; update posterior beliefs; and repeat until convergence or budget exhaustion. Acquisition functions formalize prioritization, whether through expected improvement, upper confidence bounds, Thompson sampling, or information gain. Laboratories and simulators should expose minimal, well-documented interfaces that accept proposals, return measurements with uncertainty, and log provenance for audit. When the loop is in place, learning ceases to be an offline exercise and becomes a driver of evidence, with a paper trail that links decisions to outcomes.

Reproducibility, energy awareness, and governance complete the engineering picture. A publication-grade repository includes data and model "cards" that document provenance, licenses, and intended use; environment specifications and containers that allow one-click reproduction on clean machines; seeds, hyperparameters, and scripts that regenerate tables and figures from raw inputs within a stated budget; and dashboards or logs that record energy, carbon, and cost for major training and evaluation runs. Governance addresses privacy, export controls, and dual-use risks, embedding restrictions in code and documenting access for audit. None of these requirements are ornamental; they are the conditions for work that travel across laboratories and persist beyond the original authors.

Finally, the organizational setting matters. Teams that succeed at AI for Science usually combine domain scientists, machine-learning and statistical leads, experimental engineers, and data-governance specialists. Processes that emphasize milestones, from a minimal end-to-end prototype to a closed loop, to independent verification and release, prevent premature optimization and force attention to the failure modes that papers often ignore. Blind analysis and independent replication of key claims protect against observer bias and accidental leakage [16, 17]. Common platforms for data schemas, schedulers, laboratory interfaces, and monitoring reduce bespoke glue code, making it more likely that good ideas survive contact with new data and new users.

4. Conclusion

The account developed here is intentionally plain: encode structure before learning, use learning to accelerate within the feasible set, quantify and calibrate uncertainty so that decisions can be defended, and make results reproducible and auditable so that others can build on them. When these precepts are followed, the efficiency gains often attributed to machine learning are realized without undermining the disciplines to which they are applied. In mathematics, neural operators and physics-informed neural networks demonstrate that forward and inverse problems can be approximated quickly while honoring constraints, if evaluation checks physical validity and calibration in addition to error. In high-energy physics and materials discovery, systems that treat triggers, reconstruction, and anomaly detection, or prescreening, multi-fidelity optimization, and self-driving experimentation, as parts of a single, audited loop achieve end-to-end gains that are measurable and replicable. In chemistry, pipelines that integrate representation, feasibility-aware generation, and calibrated property prediction become practical instruments rather than demonstrations. In cross-disciplinary settings, multimodal alignment, physics consistency, and probabilistic reporting make forecasting and control more conservative in the best sense of that term. Several priorities follow. The first is the continued development of equivariant and structure-aware models, including operator-learning methods whose theoretical guarantees are beginning to catch up with practice. The second is explicit

representation of cross-scale causal structure, which would make models better at explanation and more stable under intervention. Third is interface standardization between self-driving laboratories and multi-fidelity simulation to make it simpler to compose and compare closed loops. The fourth is formalization of green metrics and audits such that the role of energy and carbon action becomes a first-class consideration; efficiency should be seen as part of scientific duty. The last is social, not technical: a culture of one-click reproducibility and independent verification should be nurtured in courses and venues, with credits and incentives to match. If heeded, AI for Science will be not only more rapid but also more reliable, legible and respectful of the physical and ethical constraints within which it must labor.

References

[1] Lu L, Jin P, Karniadakis G E. Learning nonlinear operators via DeepONet. Nature Machine Intelligence, 2021, 3: 218–229.

[2] Dutta A, Das S. Not just another survey on physics-informed neural networks (PINNs): Foundations, advances, and open problems. Journal of the ACM, 2025, 37(4).

[3] Alqarafi A, Batool H, Abbas T, Janjua J I, Ramay S A, Ahmed M. Estimating uncertainty in deep learning methods and applications. In: Proceedings of the 2024 International Conference on Computer and Applications (ICCA), 2024, pp. 1–6. IEEE.

[4] Bengio Y, Lodi A, Prouvost A. Machine learning for combinatorial optimization: A methodological tour d'horizon. European Journal of Operational Research, 2021, 290(2): 405–421.

[5] Strait J D, Moran K R, Murph A C, Hyman J D, Viswanathan H S, Stauffer P H. Covariate-informed bifidelity bias correction of distributional output. SIAM/ASA Journal on Uncertainty Quantification, 2025, 13(3): 1616–1648.

[6] Hauth J. Advances in intuitive priors and scalable algorithms for Bayesian deep neural network models in scientific applications. Doctoral dissertation, 2025.

[7] Bisram R. Predicting isotopologue counts from bulk unlabeled metabolomics data. The Cooper Union for the Advancement of Science and Art, 2023.

[8] Fang J, Gentine P. Exploring optimal complexity for water stress representation in terrestrial carbon models: A hybrid machine learning model approach. Journal of Advances in Modeling Earth Systems, 2024, 16(12): e2024MS004308.

[9] Aarrestad T, et al. Fast inference of deep neural networks in FPGAs for particle physics. Frontiers in Big Data, 2021, 4: 676580.

[10] Thais S, Calafiura P, Chachamis G, DeZoort G, Duarte J, Ganguly S, Kagan M, Murnane D, Neubauer M S, Terao K. Graph neural networks in particle physics: Implementations, innovations, and challenges. arXiv preprint arXiv:2203.12852, 2022.

[11] Saxena D, Cao J. Generative adversarial networks (GANs): Challenges, solutions, and future directions. ACM Computing Surveys (CSUR), 2021, 54(3): 1–42.

[12] Wayo D D. Ensembles of graph and physics-informed machine learning for scientific modeling in materials science: A review. Archives of Computational Methods in Engineering, 2025: 1–26.

[13] Do B, Zhang R. Multi-fidelity Bayesian optimization in engineering design. arXiv preprint arXiv:2311.13050, 2023.

[14] MacLeod B P, Parlane F G L, Morrissey T D, et al. Self-driving laboratory for accelerated discovery of thin-film materials. Science Advances, 2020, 6(20): eaaz8867.

[15] Bronstein M M, Bruna J, Cohen T, Veličković P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. arXiv:2104.13478, 2021.

[16] Hertwig R, Herzog S M, Kozyreva A. Blinding to circumvent human biases: Deliberate ignorance in humans, institutions, and machines. Perspectives on Psychological Science, 2024, 19(5): 849–859.

[17] Franklin A, Laymon R. Experimentation in physics in the 20th and 21st centuries. In: Oxford Research Encyclopedia of Physics, 2024.