The AI-Enabled Scientific Frontier

Gabriel Manso[1], Emma Fu[1], Neil Thompson[1*]

[1]MIT CSAIL, MIT FutureTech.

*Corresponding author(s). E-mail(s): neil_t@mit.edu;

Abstract

As AI's capabilities improve, it is increasingly viewed as a general scientific method. But how true are these claims, and how is this changing? We assess them using 2,507 head-to-head comparisons between AI and other scientific analysis techniques across 27 disciplines from papers published between 2000 and early 2025. We find a dichotomy. Relative to traditional statistics, AI often outperforms, but at higher computational cost. Yet nearly a quarter of cases show AI to be both more expensive and worse-performing than traditional statistical techniques, a fraction that has remained stable for a decade. Relative to scientific computing, AI often *underperforms*, but at lower computational cost. Since 2020, however, AI has strengthened and now outperforms scientific computing in more than half of comparisons. These patterns suggest that AI is not a universal replacement for existing methods, but rather a valuable—and improving—part of a new AI-enabled scientific frontier.

The future of science will depend on the analytical methods researchers choose and the performance they are willing to buy with computation [1]. Some argue that AI is becoming a general scientific method [2–4], a view reinforced by bibliometric evidence of its widespread and growing use across disciplines [5]; others view it as a powerful but context-dependent technology [6–8]. Our evidence supports the latter view, while showing that AI is increasingly reshaping the scientific frontier.

Over the past six decades, advances in scientific computing have transformed how scientists interrogate nature [9–11]. Today, many anticipate another revolution, this time enabled by artificial intelligence. Deep neural networks and large foundation models already perform tasks such as protein structure prediction, medium-range weather forecasting, and broader Earth-system prediction at or beyond long-standing analytical pipelines, often with dramatically different computational requirements [12–17].

These successes have encouraged expansive claims that AI may function as a general-purpose scientific method rather than a domain-specific tool. At the same time, there are reasons for caution. Success in one scientific setting need not translate to another; theoretical and empirical work has highlighted no-free-lunch limits, failures under distribution shift, overfitting to narrow benchmarks, and the risk of illusory understanding [8, 18–20]. The computational budgets required to train frontier models have also risen steeply [21, 22], suggesting that cost concerns may also restrict AI's usage.

Scientific disciplines are having very different experiences with AI. In climate and Earth-system modeling, neural systems can rival leading baselines at far lower runtime cost [14–16]. In tabular biomedical and economic settings, by contrast, deep models often yield inconsistent gains and may require substantial tuning and compute [23–26]. This contrast echoes findings from clinical prediction, where a systematic review found no overall performance benefit of neural networks over logistic regression [27]. Domain-specific reviews and benchmarks have advanced the literature [28], but what is still missing is a cross-domain map of where AI sits in the broader cost–performance landscape. Without such a map, researchers risk overgeneralizing from their own experience.

In our analysis, we focus on AI's analytic capabilities, specifically its ability to replace other techniques from statistical analysis or scientific computing. This is distinct from work examining AI systems designed to assist or automate tasks currently performed by scientists themselves [29–31]. Our focus matters because analytical techniques often set both the quality ceiling and the computational bottleneck of scientific work. If AI systematically changes that trade-off, it changes not only scientific methods, but the pace and scope of scientific discovery itself [32, 33].

The productive question, then, is not whether AI is universally superior, but where it sits relative to existing alternatives and how that position is changing over time. To answer this, we assemble 2,507 direct comparisons from studies published between 2000 and early 2025, each reporting predictive performance and computational cost for at least two methods on the same task and dataset. We classify methods into traditional statistics, scientific computing, and AI [34], and group 27 disciplines into seven clusters based on dominant computational methods. This lets us test whether AI's effectiveness depends on what it replaces.

Historically, scientific analysis has been anchored by two method families: traditional statistics, which usually offer low computational cost and solid baseline performance, and scientific computing, which offers higher-fidelity representations of complex systems but often at far greater computational expense. We hypothesize that AI can alter this landscape in two distinct ways: by outperforming traditional statistical methods at additional computational cost, or by enabling (better or worse) scientific-computing results at far lower cost. Figure 1 illustrates this logic and introduces the reference-class structure that underlies the rest of the paper.

Five findings organize the discussion. The central one is that AI is best understood as creating a new point on the scientific frontier, rather than merely an intermediate compromise. Second, when AI replaces traditional statistical methods, the modal outcome is higher performance at higher computational cost; when AI

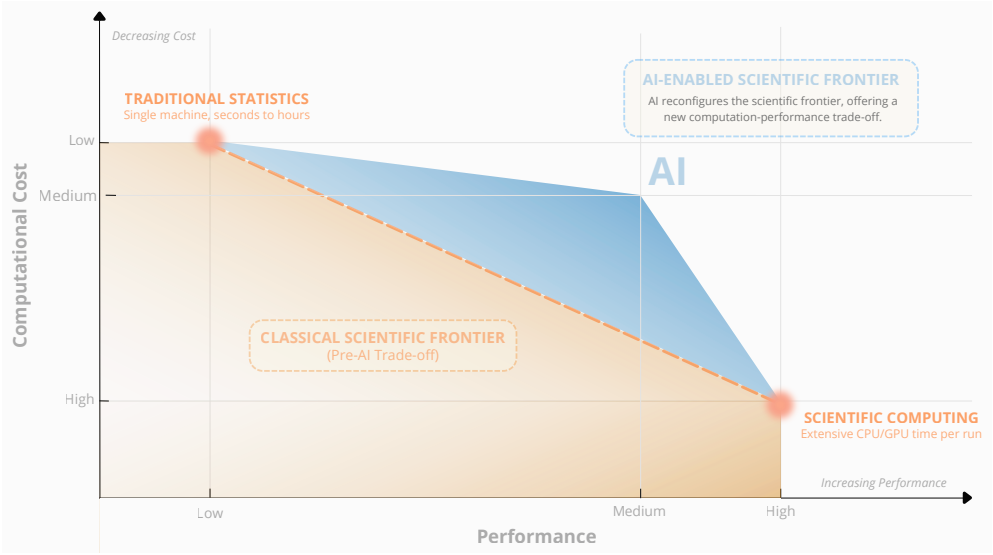


Fig. 1 The AI-enabled scientific frontier—hypothesis. Conceptual illustration of how AI can extend the historical compute–performance trade-off defined by traditional statistics and scientific computing. In this interpretation, AI is not merely an intermediate option: it opens up a new position on the frontier that offers higher performance than traditional statistics at moderate additional cost, or much lower cost than scientific computing at a reduced level of performance.

replaces scientific-computing baselines, it provides a lower cost option that sometimes improves performance and sometimes does not. Third, the impact of AI adoption is highly domain dependent. Fourth, AI's success at replacing other methods has accelerated since 2020. And, fifth, **all** scientific methods are evolving over time, and thus AI's relative position on the scientific frontier is shifting compared to other techniques.

Results

Analysis techniques across scientific domains

Whether AI looks transformative or incremental depends on the baseline. Relative to an operational weather ensemble that consumes millions of core-hours per day, modest accuracy gains at much lower runtime can be consequential. Relative to a well-regularized random forest on a small tabular dataset, the question is whether any gain justifies the added compute. Figure 2 makes this reference-class problem

concrete. For our analysis, we infer the reference class based on the comparisons made by the literature—that is, we consider a comparison technique to be relevant if a study incorporates and documents a comparison analysis as an integral element of its empirical or analytical framework.

On the statistical side, AI is usually compared not with trivial models but with strong workhorses such as support vector machines, random forests, logistic regression, decision trees, and k-nearest neighbors. On the scientific-computing side, the baselines include stochastic simulations, geophysical system models, partial differential equation solvers, analytic solutions, and quantum or atomistic simulations. These methods are often the backbone of high-fidelity scientific workflows and can be extremely computationally demanding. Figure 2a therefore shows that claims that "AI performs better" can refer to very different kinds of replacement, depending on the incumbent method family.

That heterogeneity is not incidental; it is a structural feature of the corpus. Figure 2b shows that in Earth & Space Sciences and Physical Modeling Sciences, AI is evaluated almost equally often against traditional statistical and scientific-computing baselines. By contrast, Life Sciences, Engineering, Environmental & Agricultural Sciences, and the heterogeneous "Other" cluster remain overwhelmingly statistics-heavy.

Figure 2 therefore reveals two important facts. First, AI's baselines are not uniform. In some fields it is mainly trying to displace traditional statistics, whereas in others it is competing directly with mature numerical solvers. Second, where AI is evaluated against scientific computing, the baseline methods are typically high-fidelity models rather than toy simulations, implying both substantial headroom for computational savings and high standards for fidelity.

A)

METHOD	DESCRIPTION	%
TRADITIONAL STATISTICS (n=755)		
Support Vector Machine	Kernel-based classifier that finds the maximum-margin separating hyperplane.	25.9%
Random Forest	Ensemble of decision trees using bootstrap sampling and random feature selection.	12.7%
Logistic Regression	Linear classifier using a sigmoid function for binary or multiclass prediction.	8.1%
Decision Tree	Tree-based model making sequential splits on features to classify data.	8.0%
k-Nearest Neighbors	Non-parametric method classifying points by the majority label of their k nearest neighbors.	5.2%
SCIENTIFIC COMPUTING (n=297)		
Stochastic Simulations	Models using randomness or sampling to represent uncertainty or estimate expectations.	15.5%
Geophysical System Simulations	Physics-based models of specific atmospheric, oceanic, climatic, or planetary processes.	14.1%
Partial Differential Equations Numerial Solvers	Numerical methods that discretize PDEs to compute approximate solutions.	7.7%
Closed-Form Solutions	Exact or analytic solutions used as theoretical or validation references.	6.7%
Quantum & Atomistic Simulations	Computational models of systems at quantum or atomic scales.	6.4%

B)

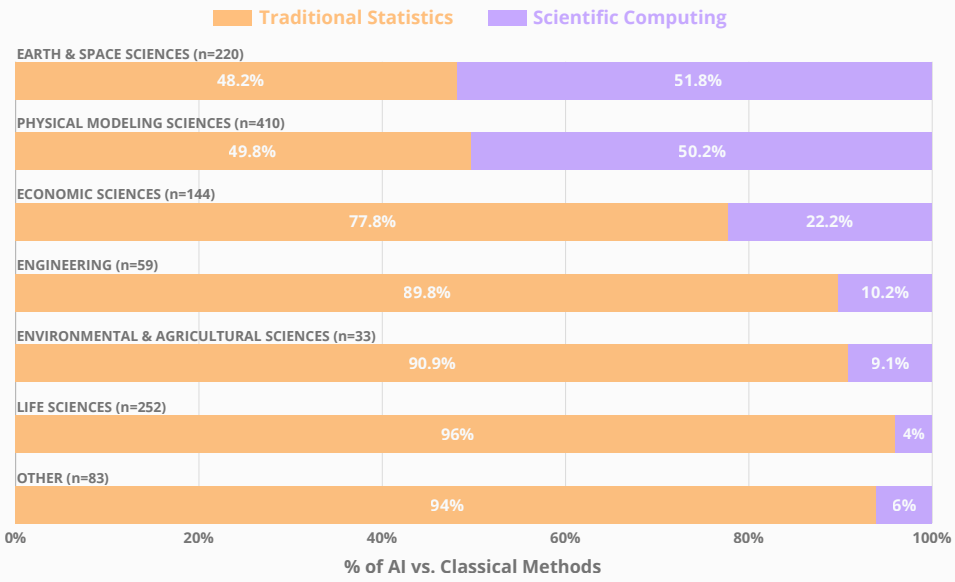


Fig. 2 AI's reference class across scientific fields. a, Most common traditional-statistics and scientific-computing baselines used in head-to-head AI comparisons. **b**, Share of AI comparisons against each baseline family across domain clusters.

What happens when AI replaces existing methods?

When researchers adopt AI in place of an existing method, what happens to performance and cost? Because these measures are reported in field-specific units (from accuracy and energy-per-atom to Courant-number violations and core-hours), we harmonize comparisons using two encoding schemes. With our first encoding scheme, we assign each replacement to a discrete outcome category based on whether AI improves or degrades *performance* and *cost*. An outcome is *win-win* when AI achieves both better performance and lower cost, *performance prioritization* when AI improves accuracy but at higher cost, *efficiency prioritization* when AI reduces cost but hurts performance, and *lose-lose* when AI is both less accurate and more expensive than the baseline it replaces. Over all the comparisons, we find that AI achieves win-win and lose-lose outcomes in nearly equal proportions (21.1% and 19.3%). Whereas, AI often allows performance prioritization (45.6%) and only sometimes efficiency prioritization (13.9%). An even clearer pattern emerges when we disaggregate these results based on whether AI is being compared to traditional statistics or scientific computing, as shown in Figure 3.

Four examples from our corpus of comparisons illustrate these scenarios:

- **Win-Win: Scientific computing → AI.** Tubiana et al. predict protein–protein binding sites using a geometric deep learning model in place of a structural homology baseline, improving AUCPR from 0.613 to 0.694 while reducing computation time from ≈30 days to ≈1.5 hours [35].
- **Performance prioritization: Traditional statistics → AI.** Davagdorj et al. predict patient risk of non-communicable diseases from national health survey data using a neural network in place of k-nearest neighbors, improving classification accuracy from 85.0% to 95.0% but at ≈10.5× longer execution time [36].
- **Efficiency prioritization: Scientific computing → AI.** George and Huerta estimate gravitational-wave parameters with a convolutional neural network (CNN)

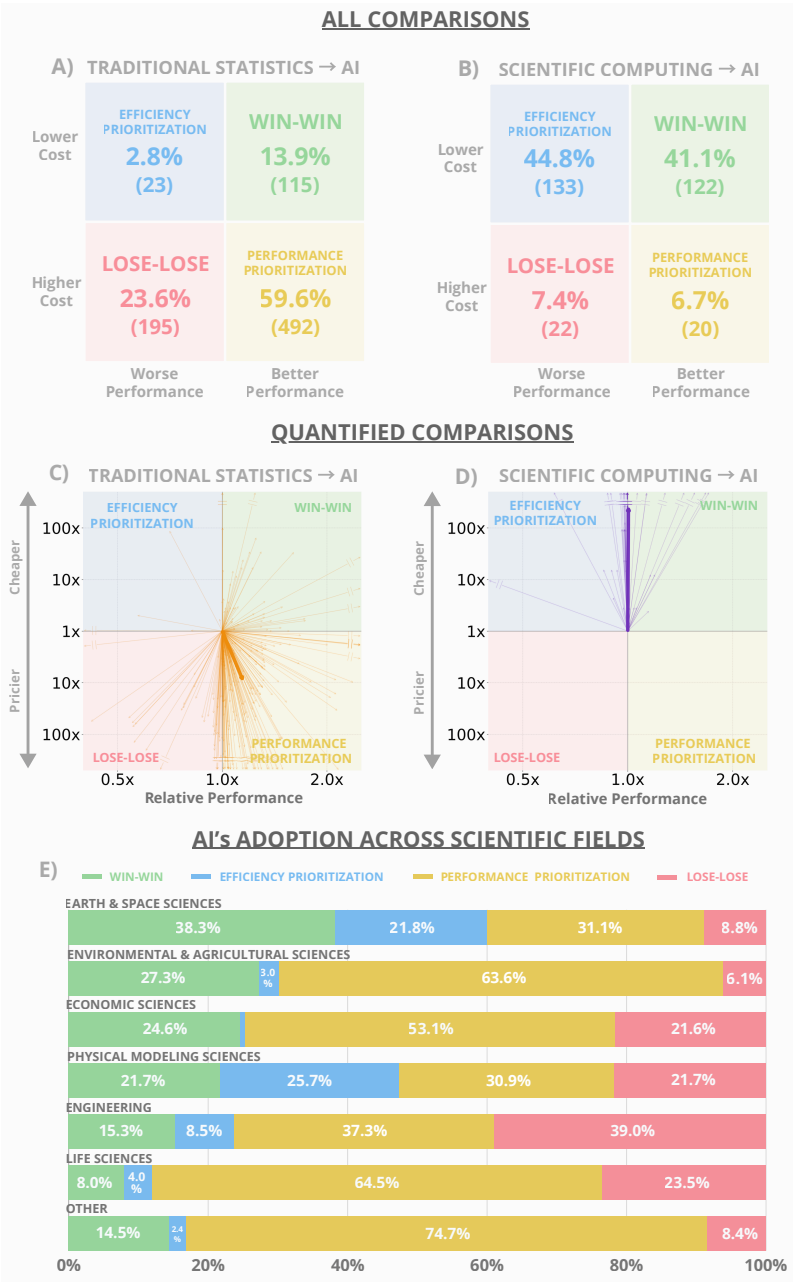


Fig. 3 AI outcomes when replacing existing methods. a,**b**, Outcome categories based on whether AI improves performance and/or computational cost relative to the baseline it replaces. **c**,**d**, Quantitative comparisons plotted in normalized performance–cost space. **e**, Outcome distributions across scientific domain clusters.

rather than a matched-filter pipeline, reducing inference time by ≈10,000× but increasing mean relative error from 18% to 22% [37].

- **Lose-Lose: Traditional statistics → AI.** Ebiwonjumi et al. predict the decay heat of spent light-water-reactor fuel assemblies with a neural network instead of Gaussian process regression, finding that mean absolute error rose from 4.28 to 5.56 while total training time increased by ≈32× [38].

When numerical values are available, as in these examples, we compute our second encoding for outcomes: log-ratio measures that capture the magnitude of change and locate each comparison in a common cost–performance plane by anchoring the baseline technique at the origin and orienting so that positive values indicate improvement. Full details of corpus construction, taxonomy definitions, metric harmonization, and robustness checks are provided in the Supplementary Information.

The results differ sharply depending on what AI replaces. When AI replaces traditional statistics (Fig. 3a), the dominant pattern is better performance at higher computational cost. In 59.6% of comparisons, AI is performance-prioritizing; 13.9% are win-win, 23.6% lose-lose, and only 2.8% efficiency-prioritizing. Quantitatively, these replacements raise performance by 23.2% on average ($\bar{\Delta}_{\text{perf}} = 0.090$, $n = 794$) while increasing computational cost by about one order of magnitude ($\bar{\Delta}_{\text{cost}} = 0.996$, or $\approx 10\times$, $n = 298$), both as geometric means. Among the $n = 293$ comparisons reporting both metrics jointly (Fig. 3c), the pattern is similar: +20.2% performance at $\approx 10\times$ higher cost. This reinforces the typical trade-off in science in which better performance can be achieved at the cost of additional computation.

When AI replaces scientific computing (Fig. 3b), the dominant pattern is lower computational cost, sometimes with and sometimes without a performance gain. AI is win-win in 41.1% of comparisons and efficiency-prioritizing in 44.8%; lose-lose (7.4%) and performance prioritization (6.7%) are uncommon. Across all comparisons with available data, AI improves performance by 26.4% on average ($\bar{\Delta}_{\text{perf}} = 0.102$, $n = 176$)

while reducing cost by about 2.8 orders of magnitude ($\bar{\Delta}_{\text{cost}} = -2.788$, or $\approx 614\times$ cheaper, $n = 137$). Among the smaller subset of $n = 69$ comparisons that report both metrics jointly (Fig. 3d), the cost reduction remains substantial ($\approx 284\times$) while performance is approximately matched ($\bar{\Delta}_{\text{perf}} = 0.002$); the difference reflects a reporting pattern in which studies demonstrating large performance gains over scientific computing baselines less frequently quantify computational cost. These results reflect AI's ability to preserve the incumbent model's fidelity while delivering near-simulation-level performance without the associated computational burden [39].

The results differ sharply depending on what AI replaces. When AI replaces traditional statistics (Fig. 3a), the dominant pattern is better performance at higher computational cost. In 59.6% of comparisons, AI is performance-prioritizing; 13.9% are win-win, 23.6% lose-lose, and only 2.8% efficiency-prioritizing. Quantitatively, these replacements raise performance by 23.2% on average ($\bar{\Delta}_{\text{perf}} = 0.090$) while increasing computational cost by about one order of magnitude ($\bar{\Delta}_{\text{cost}} = 0.996$, or $\approx 10\times$), both as geometric means (see Fig. 3c). This reinforces the typical trade-off in science in which better performance can be achieved at the cost of additional computation.

When AI replaces scientific computing (Fig. 3b), the dominant pattern is lower computational cost, sometimes with and sometimes without a performance gain. AI is win-win in 41.1% of comparisons and efficiency-prioritizing in 44.8%; lose-lose (7.4%) and performance prioritization (6.7%) are uncommon. Where quantitative comparisons are available, shown in Fig. 3d, AI improves performance by 26.4% on average ($\bar{\Delta}_{\text{perf}} = 0.102$) while reducing cost by about 2.8 orders of magnitude ($\bar{\Delta}_{\text{cost}} = -2.788$, or $\approx 614\times$ cheaper). These results reflect AI's ability to preserve the incumbent model's fidelity while delivering near-simulation-level performance without the associated computational burden [39].

AI's value therefore depends on what it replaces, but not because it usually fails against traditional statistics and only succeeds against simulation. Relative to traditional statistics, AI frequently opens a performance-for-compute trade-off. Relative to scientific computing, AI is roughly split between being a cheaper-but-imperfect surrogate and being strictly better than the comparison technique.

Domain-specific effectiveness and the migration of effort

Figure 3e shows that AI's gains are not uniform across science. Earth & Space Sciences show the clearest broad gains, with win-win outcomes accounting for 38.3% of comparisons. This is followed by Environmental & Agricultural Sciences (27.3%) and Economics (24.6%). Conversely, Life Sciences has the lowest share of win-win outcomes (8.0%) followed by Engineering (15.3%), which is similar to the Other category (14.5%). By contrast, lose-lose cases are most common in Engineering (39.0%) followed by Life Sciences (23.5%), Physical Modeling Sciences (21.7%), and Economic Sciences (21.6%).

A natural question is how much of this variation reflects genuine domain-specific differences in AI's effectiveness, as opposed to differences in what AI is being compared against. Because replacements of traditional statistics and of scientific computing have sharply different outcome profiles (Fig. 3a,b), fields with more scientific-computing baselines will mechanically produce more win-win and efficiency prioritization outcomes regardless of any domain-specific effect. To disentangle the two, we weight the corpus-wide outcome rates for each baseline family by each domain's baseline mix (Fig. 2b) to obtain expected outcome shares under the null hypothesis that AI performs identically everywhere, with variation arising solely from what it replaces. Expressed as the deviation between observed and expected rates, the domains where AI outperforms its baselines more often than composition alone would predict are Environmental & Agricultural Sciences (+21.0 pp), Other (+16.9 pp), Earth & Space

Sciences (+8.4 pp), and Economic Sciences (+4.5 pp). Life Sciences is near parity (+0.5 pp), while Physical Modeling Sciences (−9.3 pp) and Engineering (−17.0 pp) fall below expectation. On the computational side, Earth & Space Sciences (+9.3 pp) and Environmental & Agricultural Sciences (+8.6 pp) produce cheaper-than-expected AI models, whereas Life Sciences (−5.9 pp) and the remaining fields show modest negative deviations.

As these deviations show, several domains deviate substantially from the null. The cleanest test comes from Earth & Space Sciences and Physical Modeling Sciences, which share nearly identical baseline compositions (≈55% traditional statistics, ≈45% scientific computing) yet exhibit sharply divergent outcomes ($\chi^2 = 25.8$, $df = 3$, $p = 1.0 \times 10^{-5}$). Earth & Space Sciences achieves a win-win rate of 38.3%—twelve percentage points above its composition-predicted rate of 26.2%—while Physical Modeling Sciences falls below the same expected rate at 21.7%, with correspondingly elevated lose-lose outcomes (21.7% observed vs. 16.4% expected). Because baseline composition is effectively held constant in this comparison, the divergence overwhelmingly reflects domain-specific factors: differences in the structure of prediction tasks, the maturity of incumbent methods, or the intrinsic learnability of the underlying data. A second, independent line of evidence points in the same direction. Even when restricted to traditional-statistics-only comparisons—removing composition effects entirely—Life Sciences achieves a win-win rate of just 6.2%, less than half the corpus-wide traditional-statistics average of 13.9%, while its performance prioritization rate (66.9%) exceeds the same average by seven percentage points. These patterns confirm that the heterogeneity in Fig. 3e is not merely a compositional artifact. Some disciplines may feature problems whose underlying structure is intrinsically more amenable to neural approximation—for instance, spatiotemporal fields with strong physical regularities—while others may present data manifolds that are

harder to compress into the continuous, low-dimensional representations that neural architectures favour [40, 41].

Overall, AI's usage in most disciplines leads to performance prioritization, with higher costs but also higher performance. This is consistent with their wide-scale usage of baseline techniques from traditional statistics and the higher computational cost that is usually required to improve it.

How AI's advantage has changed over time

Figure 4 shows that AI's relative advantage has not been static. It has changed substantially over time, and the form of that change depends on what AI is replacing.

When AI replaces traditional statistical methods (Fig. 4a), the early evidence is limited and mixed. In the pre-2012 sample, the observed cases consist entirely of performance prioritization (58.3%) and lose-lose (41.7%) outcomes, indicating that early AI systems did not provide cheaper substitutions for traditional statistics. In 2012–2015, results are mixed: lose-lose outcomes account for 43.6% of cases, while performance prioritization and efficiency prioritization each account for 20.5%, and win-win outcomes remain limited (15.4%). After 2015, however, the pattern seems to stabilize. In 2016–2019, performance prioritization becomes the dominant outcome (56.1%), with lose-lose outcomes falling to 21.9%. By 2020–2024, this pattern strengthens further: 63.4% of substitutions are performance prioritization, whereas 22.5% are lose-lose. The mature pattern is that AI delivers higher performance at higher computational cost *and that this pattern has been stable for nearly a decade.*

The temporal pattern looks different when AI replaces scientific computing (Fig. 4b). Favourable outcomes are common from the outset but initially arise mainly through efficiency gains. Before 2012, 75.6% of cases are efficiency prioritization and 22.2% win-win. By 2016–2019, the distribution broadens but remains favourable overall (49.2% efficiency prioritization, 30.5% win-win). In 2020–2024, the balance shifts

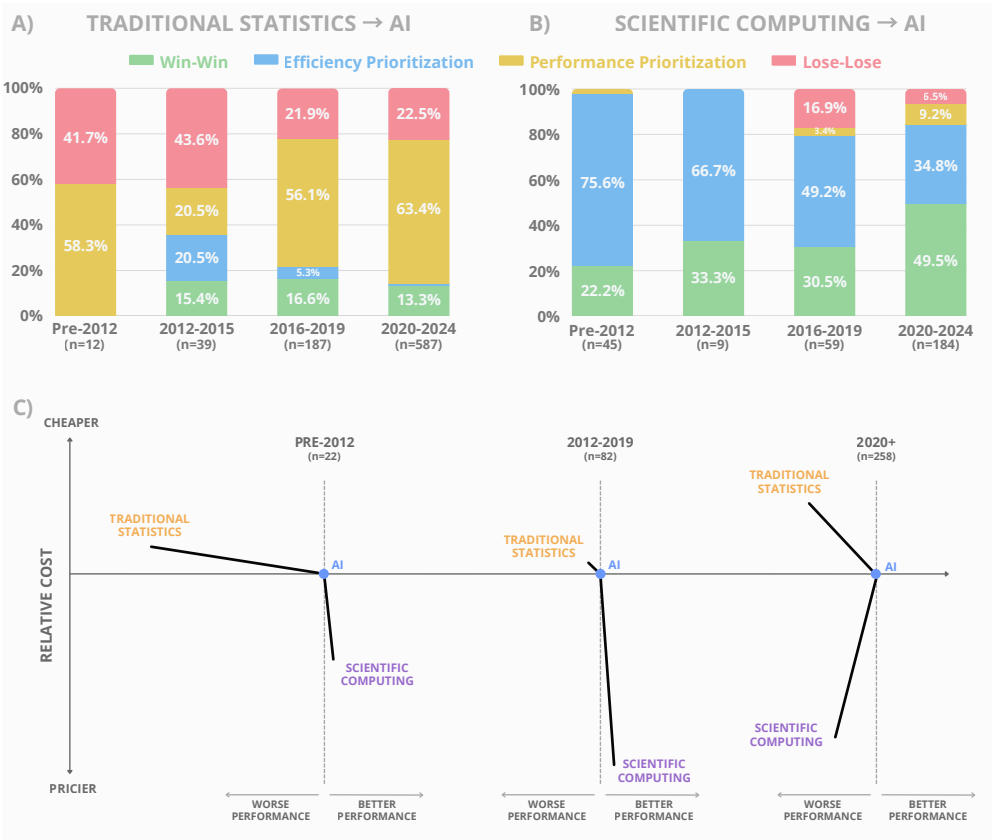


Fig. 4 How AI's advantage has changed over time. a,**b**, Outcome distributions across time when AI replaces traditional statistics and scientific computing. **c**, Average normalized positions of traditional statistics, scientific computing, and AI, showing a shift from an intermediate position to AI outperforming scientific computing on average.

toward stronger dominance: 49.5% of substitutions are win-win, 34.8% efficiency prioritization, and only 6.5% lose-lose. AI thus begins as a cheaper-but-imperfect surrogate and increasingly becomes a strict improvement on both performance and cost.

Figure 4c integrates these temporal changes by locating traditional statistics, scientific computing, and AI in the normalized cost–performance plane across broad periods. Recall that in Fig. 1, we hypothesized that AI would occupy an intermediate position on the performance–cost curve. This is indeed the pattern prior to 2020, but we can also see a new pattern emerging post 2020. More specifically, in the pre-2012 period, AI occupies an intermediate position between the two historical anchors

of scientific analysis: traditional statistics are cheaper but worse-performing, whereas scientific computing is more expensive but better-performing. This same broad ordering persists in 2012–2019, but with much smaller differences in performance between the techniques. The computational cost difference in this period between AI and scientific computing is particularly pronounced. The clearest reordering appears in the 2020+ period. Traditional statistics remain cheaper but worse-performing than AI, while scientific computing becomes both more expensive and less performant on average. In this sense, the figure does not merely show AI filling the space between existing approaches. It shows AI moving from an intermediate position to one that increasingly dominates some implementations of scientific computing while maintaining a performance advantage over traditional statistics. Here, we emphasize that it is only *some* implementations because even in the 2020+ period, 41% of AI implementations perform worse than the scientific computing method they are compared to (and thus remain well represented by Fig. 1).

The AI-enabled scientific frontier

The previous section documented how AI's relative position has shifted over time. Figure 5 instead takes a cross-sectional view, asking how the three method families compare when all observations are considered together in the cost–performance plane.

Traditional statistical methods occupy a low-cost, lower-performance region of the plane. Scientific computing lies at the opposite extreme, achieving higher performance but at computational costs $\approx 3{,}300\times$ greater. AI occupies a distinct position. Its performance is comparable to, and on average slightly exceeds, that of scientific computing, while $\approx 330\times$ cheaper. In other words, AI delivers near–simulation-level performance without simulation-level cost. Rather than simply filling the gap between traditional statistics and scientific computing, it introduces a new frontier point in the cost–performance landscape.

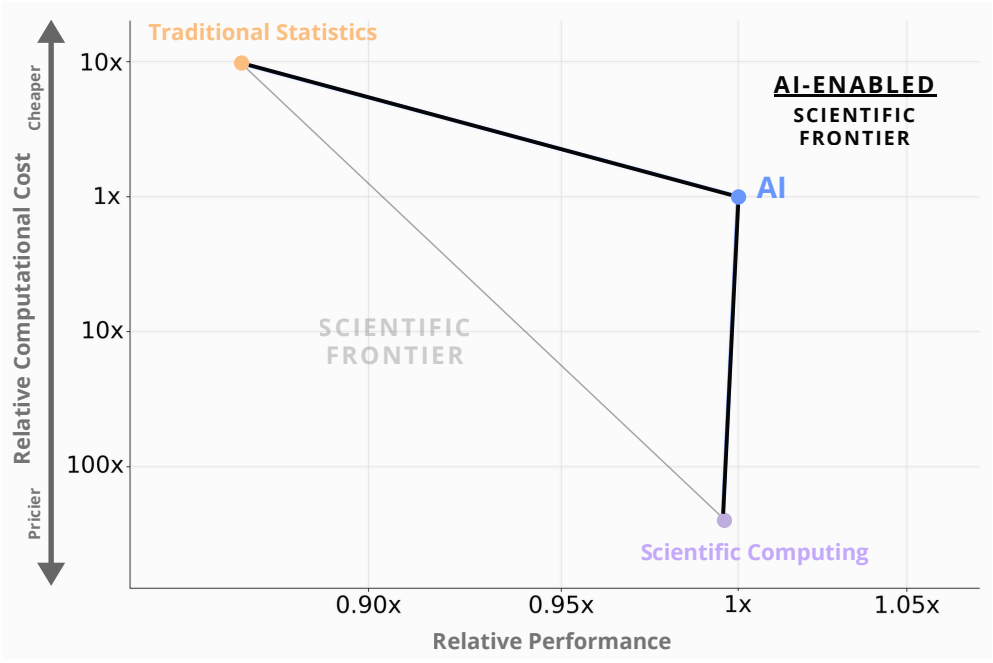


Fig. 5 The AI-enabled scientific frontier—quantified. Average positions of traditional statistics, AI, and scientific computing in normalized cost–performance space.

Figure 6 extends this perspective by adding additional head-to-head technique comparisons *within* each method family (e.g. an older AI technique compared to a newer one). It therefore shows how each method family is evolving, with arrows indicating quadrant averages for how each family is extending the scientific frontier. The picture that emerges is of an ever-deepening set of analytical techniques that offer a broad range of cost-performance options for science.

Taken together, these results clarify how the AI-enabled scientific frontier is evolving. AI does not simply replace existing approaches, nor does it uniformly dominate across the entire cost–performance space. Instead, it expands the set of attainable trade-offs in regions where simulation-based methods have historically dominated. The resulting picture is not one of wholesale replacement but of a reconfigured frontier: traditional statistics, AI, and scientific computing occupy adjoining segments of a shared

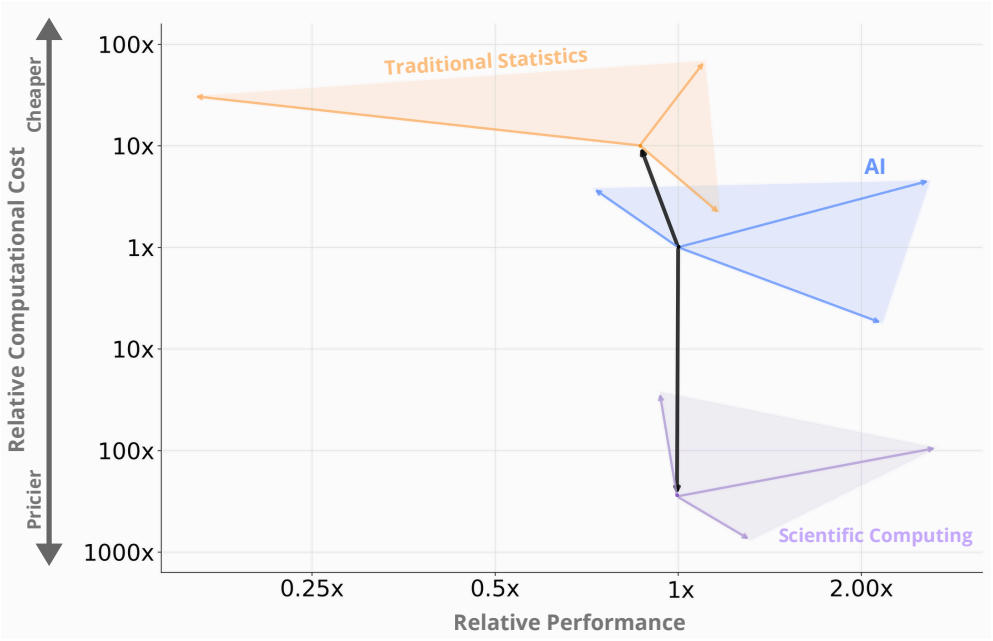


Fig. 6 Broadening regions of possibility in modern scientific methods. Distributions of observed comparisons within each method family in normalized cost–performance space.

efficiency boundary, and scientific progress increasingly comes from matching problems to the segment whose inductive biases and computational profile best suit their data and accuracy requirements.

Discussion

Our results suggest that AI is reshaping scientific analysis, but not in a single uniform way. Relative to traditional statistical methods, AI most often improves performance by drawing on additional computation. Relative to scientific computing, AI more often preserves or improves performance while sharply reducing cost. These are distinct mechanisms. One reflects scientists' willingness to spend more computation to obtain better results; the other reflects AI's growing ability to act as a high-performing surrogate for computationally intensive pipelines. Both are, at their core, patterns in how researchers choose methods under performance–cost constraints.

This distinction helps explain why AI's impact varies so strongly across fields. In simulation-heavy domains, especially Earth & Space Sciences and related physical-modeling settings, AI frequently delivers clear two-dimensional gains: lower cost with comparable or better performance. In statistics-heavy domains, the more common pattern is different. AI often improves performance, but only at higher computational cost, making its value depend on whether those gains justify the added burden of training, tuning, and deployment [23–26, 42]. Engineering remains more mixed, underscoring that AI's scientific value depends not only on model class but also on the structure of the underlying problem.

The temporal evidence suggests that these patterns are not static. Before 2020, AI often occupied an intermediate position between traditional statistics and scientific computing. In the most recent period, AI more often exceeds scientific computing on average while maintaining a performance advantage over traditional statistics. This does not establish an inexorable trend, nor does it imply that all scientific domains will move in the same direction. It does, however, indicate that AI's role in science is becoming more expansionary than marginal.

Our study has several limitations. Only about half of the comparisons report numerical cost metrics, and those costs are measured in heterogeneous units; log-ratio encodings and robustness checks help, but cannot eliminate reporting bias. The corpus also overrepresents fields with stronger benchmarking and baseline-reporting norms, limiting generalization. Finally, our method taxonomy and domain clusters necessarily aggregate heterogeneous approaches, and we do not examine other dimensions of scientific value such as interpretability, calibration, robustness, or hypothesis generation.

These limitations suggest three priorities for future work: standardized reporting of computational budgets and hardware for all methods, including baselines [43, 44];

richer datasets on hybrid workflows in which AI augments rather than replaces numerical pipelines; and predictive models linking problem characteristics to the relative advantage of different computational approaches.

Taken together, our findings support a more precise view of AI in science than either triumphalist or sceptical accounts allow. AI is neither a universal replacement for existing methods nor merely another tool within an unchanged landscape. It is a method family that increasingly opens new regions of the scientific cost–performance frontier: by purchasing performance relative to traditional statistics, and by reducing cost relative to simulation-heavy incumbents. The central question, therefore, is not whether AI will replace existing approaches everywhere, but where its inductive biases, computational profile, and surrogate capabilities allow it to create genuinely new possibilities for scientific analysis.

Methods

Corpus construction and literature search

We assembled a corpus of published studies that reported direct same-task comparisons between computational methods in scientific applications. The central inclusion principle was comparability. Each retained study had to evaluate at least two methods on a shared problem, dataset, or experimental setting, thereby permitting a meaningful assessment of relative performance and computational cost. By restricting the corpus to explicit head-to-head evaluations within the same task environment, we isolate the empirical contexts in which AI is not merely proposed, but directly tested against incumbent computational approaches.

Candidate papers were identified through systematic queries of five databases: IEEE Xplore, ACM Digital Library, arXiv, Google Scholar, and Scopus. Each query combined three components: domain anchor terms, AI method terms, and comparison terms. For each of the 27 disciplines, two to five query variants were constructed to

capture differences in field-specific vocabulary while preserving a common search logic. Representative query examples are provided in Supplementary Table S1.

Inclusion and exclusion criteria

A paper entered the screened corpus ($N = 880$) if it satisfied two core criteria. First, it had to present at least one explicit comparison between two or more computational methods on the same task. Second, the paper had to address a scientific or scientific-computing task and be available through a peer-reviewed venue or recognized preprint server.

Papers were promoted to the final dataset ($N = 268$ papers; $N = 2{,}507$ comparisons) if they additionally reported at least one performance metric for both methods and at least one computational cost indicator, whether quantitative or qualitative.

Papers were excluded if the comparison was purely theoretical, if only a single method was evaluated, if the task was outside the scope of scientific analysis, if the reported evidence was insufficient even for directional classification, or if the paper was a review rather than an original empirical study.

Data extraction protocol

The canonical unit of observation is one method pair evaluated on one dataset or experimental setting. This definition prevents metric multiplication within the same experiment while preserving meaningful variation across datasets, benchmarks, or environmental conditions. When a paper reported multiple metrics for the same comparison, the metric most standard to the relevant domain was selected as the primary performance indicator. The same pair of methods evaluated on distinct datasets or settings was treated as separate observations.

For each comparison, the extraction template recorded paper metadata, domain classification, baseline and proposed method names, method-family assignments, performance metric name and values, cost metric name and values, directional labels for

performance and cost, the source of any qualitative cost judgment, and free-text notes for edge cases.

To maximize consistency, information was extracted according to a fixed evidence hierarchy. Highest priority was given to tables containing numerical values, followed by explicit numerical statements in the main text, then figures, and finally qualitative author statements used only for directional classification when numerical cost data were unavailable.

LLM-assisted validation

To reduce extraction errors at scale while avoiding automation bias, we implemented a *manual-first, LLM-second* validation pipeline. All comparisons were initially identified and extracted by human reviewers. The LLM system (GPT-4) was used only as a secondary verification layer and never introduced new observations or replaced human judgments without explicit review.

The validation process proceeded in six phases. First, the model mapped the structure of each paper, identifying relevant tables, figures, and result sections. Second, it enumerated candidate method comparisons within those sections. Third, it extracted the supporting evidence associated with each comparison, including reported metrics and contextual statements. Fourth, numerical values were cross-checked against the original source to verify consistency with the recorded entries. Fifth, the model classified whether its interpretation agreed with the human extraction. Finally, any disagreements were reviewed and adjudicated by a human annotator before the comparison was retained in the dataset.

Method taxonomy

All computational methods in the corpus were classified into one of three families, corresponding to the broad methodological categories discussed in the main text and positioned along the cost–performance frontier in Figure 1.

- **Traditional statistics (TS).** Algorithms that extract patterns from empirical data without relying on neural network architectures. Representative examples include support vector machines, random forests, gradient boosting, logistic regression, k-nearest neighbors, Gaussian processes, and principal components methods. These approaches are often computationally lighter than modern neural models, although performance depends strongly on the fit between their inductive biases and the underlying problem structure.
- **Scientific computing (SC).** Methods whose algorithmic structure is derived from embedded mathematical, physical, or mechanistic principles. Examples include Navier–Stokes solvers, finite element methods, molecular dynamics, computational fluid dynamics simulations, and quantum chemistry calculations. These approaches are often computationally demanding but are typically grounded in explicit model structure, constraints, or convergence properties.
- **AI.** Methods whose primary representational mechanism is a neural network architecture. Examples include convolutional, recurrent, and transformer networks, graph neural networks, physics-informed neural networks, neural operators, and related generative or surrogate-model architectures. These methods typically rely on large numbers of learned parameters optimized through gradient-based training.

Replacement versus augmentation

Beyond the family labels, cross-family AI comparisons were further classified according to whether AI replaced an incumbent method or augmented it.

- **Replacement.** The proposed AI method operates independently and is evaluated as a substitute for the incumbent method.
- **Augmentation.** The incumbent computational pipeline remains in place, but an AI component is introduced to improve or accelerate one part of the workflow.

This distinction is important for interpretation. Replacement cases align most directly with the main paper's central question of whether AI occupies a different position on the cost–performance frontier than the methods it displaces, whereas augmentation cases more often reflect targeted optimization within an existing pipeline.

Of the 2,507 comparisons in the full dataset, 1,201 are cross-family AI comparisons. Among these, 1,122 (93.4%) are classified as *replacement* and 79 (6.6%) as *augmentation*. The replacement set comprises 825 TS→AI comparisons and 297 SC→AI comparisons. The remaining observations are within-family comparisons or other non-AI cross-family substitutions. The full distribution across all comparison types is reported in Supplementary Table S2.

Domain consolidation

To support cross-domain comparison without collapsing all scientific areas into a single undifferentiated group, we consolidated the 27 fine-grained disciplines into seven broader clusters. This mapping preserved meaningful domain structure while aligning fields that share similar computational tasks, data modalities, or methodological baselines. The resulting clustering was used throughout the descriptive analyses reported in the main text and supplement. The complete mapping, together with the defining characteristics of each cluster and the number of observations it contains, is reported in Supplementary Table S3.

Metric harmonization

The dataset spans a highly heterogeneous measurement landscape, with more than 284 distinct performance metrics and 239 distinct computational cost metrics reported across 27 scientific domains. Because these metrics differ widely in scale, interpretation, and reporting convention, direct aggregation across studies is not generally meaningful.

To preserve comparability while retaining information about the direction and, when available, the magnitude of change, we employ two complementary encoding systems.

Binary encoding.

For every comparison, both performance and cost are assigned directional labels:

$$\text{Performance binary} = \begin{cases} +1 & \text{new method achieves better predictive performance} \\ 0 & \text{equivalent within a negligible margin} \\ -1 & \text{new method achieves worse predictive performance} \end{cases} \tag{1}$$

$$\text{Cost binary} = \begin{cases} +1 & \text{new method requires greater computational resources} \\ 0 & \text{equivalent} \\ -1 & \text{new method requires fewer computational resources} \end{cases} \tag{2}$$

This directional encoding covers all 2,507 comparisons. To ensure consistency, we validated all qualitative binary labels against their corresponding quantitative log-ratios. For 'lower-is-better' performance cases, where a numerical decrease denotes a performance improvement, we inverted the corresponding log-ratios to allow for direct comparison.

Log-ratio encoding.

When numerical values are available, we compute oriented log-ratio encodings:

$$\Delta_{\text{perf}}^{X:Y} = \log_{10}\left(\frac{P_X}{P_Y}\right), \qquad \Delta_{\text{cost}}^{X:Y} = \log_{10}\left(\frac{C_X}{C_Y}\right) \tag{3}$$

where Y denotes the baseline method, X the proposed method, P a performance metric oriented so that positive values indicate improvement, and C a computational cost metric oriented so that positive values indicate increased cost.

Outcome categories

Each comparison is also assigned to one of four mutually exclusive quadrants defined by the joint binary performance and cost outcomes:

- **Win-Win (WW).** Performance is equal or better and cost is equal or lower.
- **Performance Prioritization (PP).** Performance improves, but cost increases.
- **Efficiency Prioritization (EP).** Performance worsens, but cost decreases or remains unchanged.
- **Lose-Lose (LL).** Performance worsens or remains unchanged while cost increases.

Ties (binary = 0) are grouped with favorable outcomes where they arise: performance ties with cost reduction ($N = 36$) count as Win-Win; performance ties with cost increases ($N = 16$) count as Lose-Lose; joint ties (0/0; $N = 2$) count as Win-Win.

Data completeness and selective reporting

Because the corpus draws on heterogeneous literatures with different reporting conventions, the availability of quantitative information varies across comparisons. Predictive performance metrics are usually reported numerically, whereas computational cost is often omitted or described only qualitatively. Overall completeness rates by metric type and cluster-level cost reporting rates are reported in Supplementary Tables S4 and S5. Analyses based on quantitative cost ratios necessarily draw on a more restricted subset of the literature than the directional analysis emphasized in the main text.

A potential concern in compiled datasets where computational cost is not universally reported is selective reporting. In this context, authors might be more likely

to report quantitative cost metrics when the proposed AI method appears computationally favorable, while omitting such values when models impose substantial computational burdens. To assess this possibility, we tested whether the availability of quantitative cost data depends on the reported performance outcome of the proposed AI method. Using the AI replacement sample (N = 1,122), we conducted a chi-squared test of independence between AI performance outcome (superior versus inferior/equivalent) and the presence of numerical cost data. The test yields $\chi^2 = 0.055$ (df = 1, $p = 0.814$), providing no evidence that the availability of quantitative cost measurements depends on the reported performance outcome of the proposed method. This does not eliminate all possible reporting biases, but suggests that the quantitative cost subsample is unlikely to be systematically biased along this particular dimension.

Statistical analysis

Analyses focus primarily on the subset of comparisons classified as *AI replacement* (N = 1,122), isolating settings in which a neural network–based method is evaluated as a direct substitute for an incumbent computational approach belonging to either Traditional Statistics or Scientific Computing. Restricting attention to this subset removes hybrid augmentation cases and allows the empirical structure of direct methodological substitution to be examined cleanly.

Proportions of outcome categories are reported with 95% Wilson score confidence intervals, which provide more reliable coverage for bounded proportions than normal approximations. Association between categorical variables (e.g., domain cluster and outcome quadrant) was assessed using Pearson's chi-squared test of independence, with effect magnitude summarized by Cramér's V. Quantitative effect sizes are reported using the log-ratio encodings in equation (3) and summarized by their means and medians within stratified subsamples. Supporting tables reporting outcome proportions with confidence intervals, stratification by baseline type, domain-cluster contingency

structure, and quantitative effect-size summaries are provided in Supplementary Section 5.

Robustness analyses

Because the dataset aggregates comparisons from heterogeneous literatures, we verified that the reported patterns are not driven disproportionately by domain composition or by differences in reporting convention through two robustness checks on the AI replacement sample.

First, we performed a leave-one-cluster-out sensitivity analysis, recomputing the outcome distribution after sequentially removing each of the seven domain clusters from the AI replacement sample. This addresses the possibility that a single domain cluster could dominate the aggregate results, which is particularly relevant for simulation-heavy areas such as Physical Modeling Sciences that contain many comparisons involving computationally intensive mechanistic baselines.

Second, we repeated the quadrant analysis using only the subset of comparisons for which numerical cost values are available. This addresses the concern that some studies report computational cost only qualitatively and that such reports may be less comparable than explicit numerical measurements.

Both sensitivity tables are reported in Supplementary Section 6.

Supplementary information. Supplementary Information is available for this paper, Sections S1–S6 and Tables S1–S12.

Acknowledgements. We acknowledge support from MIT's UROP Program and the contributions of the master's students who assisted with early-stage data collection: Caroline C. Warren and Haley Nakamura.

Code & Data availability. The full dataset and analysis code are available in the GitHub repository.

References

[1] Anonymous, Anonymous & Anonymous. The importance of (exponentially more) computing power (2022). URL https://arxiv.org/abs/2206.14007. arXiv:2206.14007.

[2] Bianchini, S., Müller, M. & Pelletier, P. Artificial intelligence in science: An emerging general method of invention. *Research Policy* **51**, 104604 (2022).

[3] The Royal Society. Science in the age of ai: How artificial intelligence is changing the nature and method of scientific research. Tech. Rep., The Royal Society, London, UK (2024). URL https://royalsociety.org/-/media/policy/projects/science-in-the-age-of-ai/science-in-the-age-of-ai-report.pdf.

[4] Wang, H. *et al.* Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).

[5] Gao, J. & Wang, D. Quantifying the use and potential benefits of artificial intelligence in scientific research. *Nature Human Behaviour* **8**, 2281–2292 (2024).

[6] Narayanan, A. & Kapoor, S. Ai as normal technology: An alternative to the vision of ai as a potential superintelligence. Knight First Amendment Institute essay (2025). URL https://knightcolumbia.org/content/ai-as-normal-technology.

[7] Narayanan, A. & Kapoor, S. Why an overreliance on ai-driven modelling is bad for science. *Nature* **640**, 312–314 (2025). URL https://www.nature.com/articles/d41586-025-01067-2.

[8] Messeri, L. & Crockett, M. J. Artificial intelligence and illusions of understanding in scientific research. *Nature* **627**, 49–58 (2024).

[9] Cooley, J. W. & Tukey, J. W. An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation* **19**, 297–301 (1965).

[10] Clough, R. W. *The finite element method in plane stress analysis. Proceedings of the 2nd ASCE Conference on Electronic Computation*, 345–378 (American Society of Civil Engineers, Pittsburgh, PA, 1960).

[11] Rahman, A. Correlations in the motion of atoms in liquid argon. *Physical Review* **136**, A405–A411 (1964).

[12] Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). URL https://doi.org/10.1038/s41586-021-03819-2.

[13] Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature* **630**, 493–500 (2024).

[14] Lam, R. *et al.* Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023).

[15] Price, I. *et al.* Probabilistic weather forecasting with machine learning. *Nature* **637**, 84–90 (2025).

[16] Bodnar, C. *et al.* A foundation model for the earth system. *Nature* **641**, 1180–1187 (2025).

[17] Reichstein, M. *et al.* Deep learning and process understanding for data-driven earth system science. *Nature* **566**, 195–204 (2019). URL https://doi.org/10.1038/s41586-019-0912-1.

[18] Wolpert, D. H. & Macready, W. G. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation* **1**, 67–82 (1997).

[19] Koh, P. W. *et al.* *WILDS: A benchmark of in-the-wild distribution shifts. Proceedings of the 38th International Conference on Machine Learning*, Vol. 139 of *Proceedings of Machine Learning Research*, 5637–5664 (PMLR, 2021). URL https://proceedings.mlr.press/v139/koh21a.html.

[20] Liao, T., Taori, R., Raji, I. D. & Schmidt, L. *Are we learning yet? a meta review of evaluation failures across machine learning. Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* (2021). URL https://openreview.net/forum?id=mPducS1MsEK.

[21] Cottier, B. *et al.* The rising costs of training frontier ai models. *arXiv preprint arXiv:2405.21015* (2024). URL https://arxiv.org/abs/2405.21015.

[22] Anonymous, Anonymous, Anonymous & Anonymous. *The computational limits of deep learning. Proceedings of the 2023 ACM Conference on Computing and Sustainable Societies (LIMITS '23)* (ACM, 2023). URL https://doi.org/10.21428/bf6fb269.1f033948.

[23] Shwartz-Ziv, R. & Armon, A. Tabular data: Deep learning is not all you need. *Information Fusion* **81**, 84–90 (2022).

[24] Grinsztajn, L., Oyallon, E. & Varoquaux, G. Why do tree-based models still outperform deep learning on tabular data? *arXiv preprint arXiv:2207.08815* (2022). URL https://arxiv.org/abs/2207.08815.

[25] Makridakis, S., Spiliotis, E. & Assimakopoulos, V. Statistical and machine learning forecasting methods: Concerns and ways forward. *PLOS ONE* **13**, e0194889 (2018).

[26] Makridakis, S. *et al.* Statistical, machine learning and deep learning forecasting methods: Comparisons and ways forward. *Journal of the Operational Research*

Society **74**, 840–859 (2023).

[27] Christodoulou, E. *et al.* A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology* **110**, 12–22 (2019). URL https://doi.org/10.1016/j.jclinepi.2019.02.004.

[28] Thiyagalingam, J., Shankar, M., Fox, G. & Hey, T. Scientific machine learning benchmarks. *Nature Reviews Physics* **4**, 413–420 (2022).

[29] Karpathy, A. autoresearch (2026). URL https://github.com/karpathy/autoresearch. GitHub repository.

[30] Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).

[31] Szymanski, N. J. *et al.* An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature* **624**, 86–91 (2023).

[32] Besiroglu, T., Emery-Xu, N. & Anonymous. Economic impacts of ai-augmented rd. *Research Policy* **53**, 105037 (2024). URL https://www.sciencedirect.com/science/article/pii/S0048733324000866.

[33] Jones, B. F. Artificial intelligence in research and development. Working Paper 34312, National Bureau of Economic Research (2025).

[34] Breiman, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science* **16**, 199–231 (2001). URL https://doi.org/10.1214/ss/1009213726.

[35] Tubiana, J., Schneidman-Duhovny, D. & Wolfson, H. J. ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction.

Nature Methods **19**, 730–739 (2022).

[36] Davagdorj, K., Bae, J.-W., Pham, V.-H., Theera-Umpon, N. & Ryu, K. H. Explainable artificial intelligence based framework for non-communicable diseases prediction. *IEEE Access* **9**, 123672–123688 (2021).

[37] George, D. & Huerta, E. A. Deep neural networks to enable real-time multimessenger astrophysics. *Physical Review D* **97**, 044039 (2018).

[38] Ebiwonjumi, B., Cherezov, A., Dzianisau, S. & Lee, D. Machine learning of LWR spent nuclear fuel assembly decay heat measurements. *Nuclear Engineering and Technology* **53**, 3563–3579 (2021).

[39] Azizzadenesheli, K. *et al.* Neural operators for accelerating scientific simulations and design. *Nature Reviews Physics* **6**, 320–328 (2024). URL https://doi.org/10.1038/s42254-024-00712-5.

[40] Pope, P., Zhu, C., Abdelkader, A., Goldblum, M. & Goldstein, T. *The intrinsic dimension of images and its impact on learning. International Conference on Learning Representations (ICLR)* (2021). URL https://openreview.net/forum?id=XJk19XzGq2J.

[41] Ansuini, A., Laio, A., Macke, J. H. & Zoccolan, D. *Intrinsic dimension of data representations in deep neural networks. Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 32 (2019). URL https://arxiv.org/abs/1905.12784.

[42] Sculley, D. *et al.* *Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems*, Vol. 28, 2503–2511 (2015). URL https://proceedings.neurips.cc/paper/5656-hidden-technical-debt-in-machine-learning-sy.

[43] Schwartz, R., Dodge, J., Smith, N. A. & Etzioni, O. Green ai. *Communications of the ACM* **63**, 54–63 (2020).

[44] Henderson, P. *et al.* Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research* **21**, 1–43 (2020). URL https://jmlr.org/papers/v21/20-312.html.