OPEN ACCESS

Artificial Intelligence COMMENTARY

When AI meets physics: Beyond prediction, towards understanding

Jun Li,[1,*] Ruichen Ren,[1] Zhixuan Zhao,[1] and Yanling Wu[1]

[1]State Key Laboratory of Metastable Materials Science and Technology & Hebei Key Laboratory of Microstructural Material Physics, School of Science, Yanshan University, Qinhuangdao 066004, China

*Correspondence: ljcj007@ysu.edu.cn

Received: May 4, 2026; Accepted: July 30, 2026; Published Online: July 30, 2026; https://doi.org/10.59717/j.tip.2026.100006

Citation: Li J., Ren R.C ., Zhao Z.X., et al. (2026). When AI meets physics: Beyond prediction, towards understanding. *The Innovation Physics* **1**:100006.

Recent years have witnessed a rapid convergence between artificial intelligence (AI) and physics. Across quantum many-body theory, molecular simulation, fluid mechanics, astrophysics, high-energy collisions, and materials discovery, machine learning methods now achieve predictive performance with astonishing accuracy. In many subfields, AI is no longer peripheral support, but an increasingly visible component of the research workflow.[1] Yet a foundational question persists: when AI makes a correct prediction, have we truly advanced physical understanding?

Physics is distinguished from pure engineering by its pursuit of concise, interpretable laws, not merely reproducible outputs. Although accurate prediction has often opened the path toward deeper explanation, it is not sufficient on its own. If AI's predictions cannot be translated into physical insight, the current wave of "AI for physics" risks using computational power to mask ignorance rather than illuminate nature. We argue that genuine integration of AI into physics requires moving beyond a purely black-box prediction mindset. Interpretability should be elevated to a core evaluation criterion, and physical priors should be embedded directly into AI design where the scientific task demands them. Only then can AI move from "getting the numbers right" to "making sense".

WHAT DOES INTERPRETABILITY MEAN IN PHYSICS?

Before judging whether an AI model is interpretable, we should clarify what an "understanding" means to a physicist. This meaning is not identical across all branches of physics: some fields emphasise compact equations, others emphasise symmetry principles, causal mechanisms, reproducible simulations, or experimentally testable latent variables. Computer science provides post-hoc tools such as Layer-wise Relevance Propagation (LRP), SHapley Additive exPlanations (SHAP), and attention maps. These methods can reveal which inputs or internal components influence a prediction, and they have made meaningful contributions in many scientific applications. They are therefore useful diagnostics, but not always sufficient as physical explanations.

For a model to be considered physically interpretable, we propose that it should satisfy three conditions (Figure 1).

- First, it should reveal hidden physical quantities, like symmetries, conserved quantities, order parameters, or effective degrees of freedom. A neural network that classifies phases but cannot identify the order parameter or symmetry-breaking pattern may still be useful, but it adds limited new physical knowledge.
- Second, it should respect known physical laws. The model's logic should be compatible with energy conservation, causality, locality, etc. If a model makes correct predictions but violates these principles (e.g., using non-local correlations for a local system), its success should not yet be treated as physical insight.

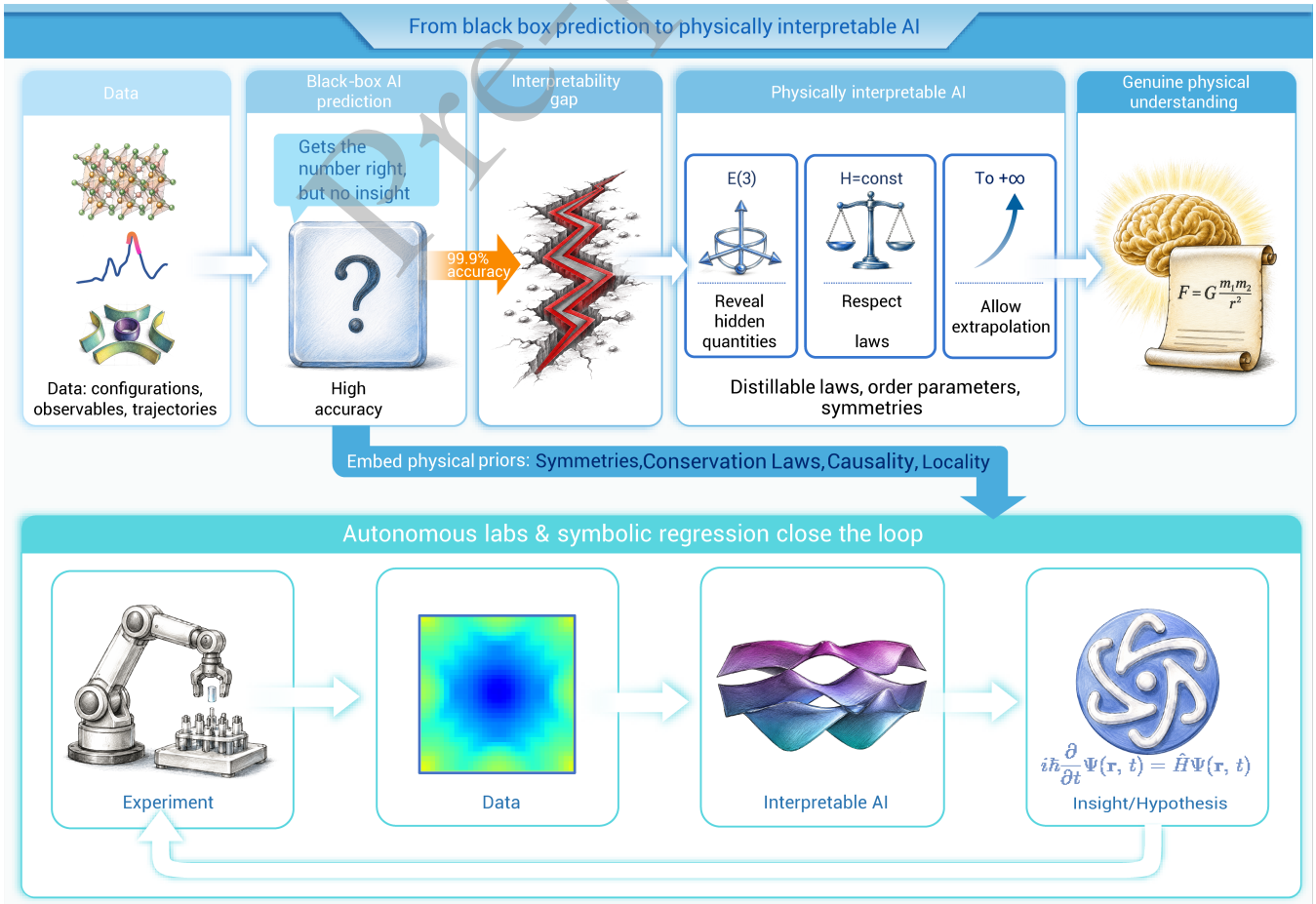

Figure 1. From prediction to understanding in AI-driven physics Black-box models achieve high accuracy but lack insight. Embedding physical priors enables interpretable AI that reveals laws, respects constraints, and supports extrapolation.

● Third, it should allow theoretical extrapolation and generalization. An interpretable model should ideally be distillable into a concise mathematical form that enables predictions for unseen parameters or scales; otherwise it remains close to a high-dimensional lookup table.

By these standards, most current explainable AI techniques offer local, descriptive accounts rather than principled, structural explanations. This gap is the central dilemma of AI for physics: we have powerful predictors, but far fewer systems that yield new physical theories.

DIAGNOSING THE INTERPRETABILITY DILEMMA

Measured against the above criteria, many celebrated AI applications in physics exhibit the same pattern: predictive power often advances faster than interpretability (upper panel of Figure 1).

In quantum many-body physics, neural quantum states can represent strongly correlated ground states with impressive accuracy. Yet the resulting variational wavefunction is typically encoded in millions of non-linear, redundant parameters. Physicists are often unable to extract transparent structures such as analytical entanglement spectra, topological correlations, or symmetry-protection mechanisms. The network "gets the numbers right" but does not easily tell us why the state takes that form.

In phase transition identification, machine learning can pinpoint transition points from raw configurations and even flag unknown phases. However, outputting a critical parameter value is not equivalent to providing a mechanism. A clustering-based model may separate phases, but it may not reveal how symmetry is broken, what scaling behavior emerges, or whether an effective field theory exists. We learn that something changes near a given parameter, but still need to know what changes and why.

These examples expose a deeper epistemological risk. When AI predictions surpass existing human intuition, it becomes harder to distinguish hints of new physics from overfitting artifacts. At the same time, even a partially interpretable AI system can guide experiments, expose anomalies, and suggest new hypotheses. Without stronger interpretability standards, however, AI risks remaining mainly an auxiliary tool for confirming what we already know and accelerating computation, rather than a pioneering partner that reveals the unknown and inspires new theories.

MAKING AI TRULY INTERPRETABLE IN PHYSICS

A natural response is to improve post-hoc explanation techniques. The more promising way out is a paradigm shift: physicists should move from passively explaining AI to actively designing AI with built-in physics priors. We propose three mutually reinforcing directions.

Embed physics laws into architecture, not just post-hoc explanation. Rather than struggling to interpret an arbitrary black box after training, physical knowledge should be encoded into the model architecture from the start when the problem structure allows it. This point is distinct from the fact that modern AI models may have been trained on large amounts of text containing physical concepts. Exposure to physical knowledge during training is not the same as enforcing conservation laws, symmetries, or limiting behaviours at the architectural, algorithmic, or evaluation level. Equivariant neural networks (E(3)NN) exemplify this principle: by constructing layers that respect symmetries such as translations, rotations, and permutations, the outputs automatically obey system invariances.[2] Internal representations can then be more directly related to observable quantities like forces and tensor fields, making interpretability closer to an inherent design property (Figure 1).

Use AI to discover physical quantities, not just fit data. The highest form of interpretability is to automatically extract physically meaningful primitives from observational data. Recent work has demonstrated symbolic optimization methods have begun to derive explicit physical laws, such as Kepler's planetary motion and gravitational wave power formulas, by integrating experimental data with axiomatic background theory and applying rigorous certificates of correctness.[3] Such methods produce mathematical structures that physicists can directly inspect and theorize about. In this mode, AI shifts from being an opaque prophet to a theorist's assistant that crystallizes latent physical structure.

Establish interpretability as an evaluation criterion for AI in physics. Currently, prediction accuracy is often the dominant metric and the model with the lowest test error is often implicitly assumed to be the best. This metric alone can be misaligned with the broader goals of physics. We call for multi-dimensional evaluation: a model's value should be judged not only by how accurately it predicts, but also by how much it helps distil new physical principles. A practical checklist could include conservation-law tests, symmetry-ablation tests, extrapolation benchmarks, failure-mode analysis, comparison with known limiting theories, and attempts to map learned variables to physical quantities. Reviewers and funders should ask: Do internal states correspond to known or new physical quantities? Do prediction failures reveal the boundaries of current theory? Can the learned structure be reduced to an analytical form? Incorporating such interpretability metrics would fundamentally shift incentives in the field.

These paths are not fantasies, frontier work demonstrates their feasibility. However they remain peripheral, overwhelmed by studies that celebrate black-box prediction accuracy. Mainstream adoption requires sustained collaboration among physicists, computer scientists, and journal editors.

EMERGING FRONTIERS AND OUTLOOK

Recent developments further reinforce the need for interpretability and open new possibilities. MatterGen, a generative model that produces stable and diverse inorganic materials while satisfying broad property constraints, has demonstrated experimental validation and advances the vision of foundational AI for materials design.[4] Autonomous laboratories that combine AI-driven experimental design with robotic execution are closing the loop between hypothesis, measurement, and model refinement, yet this loop will only generate lasting physical insight if the models are interpretable.[5] Moreover, the integration of symbolic regression with deep learning holds the promise of automatically recovering Hamiltonian, Lagrangian, or conservation law formulations.

Looking ahead, the symbiosis of AI and physics demands that interpretability becomes a first-class design objective, not an afterthought. We envision a future in which AI model deployed in physics carries a physically meaningful explanation route, and where review standards reward depth of insight just as much as benchmark performance. Only by holding AI to the same standards of understanding that physics has cherished for centuries can we ensure that the upcoming flood of AI-driven discoveries constitutes genuine progress rather than statistical illusions.

FUTURE TRENDS

AI and physics stand at a crossroads. One road continues to treat AI as a powerful prediction tool, content with black-box success, smooth and crowded but risking a technicist dead end. The other road weaves AI into physics' theoretical tradition, using interpretability as the bridge to genuine understanding, rugged and less travelled but leading to deep symbiosis. Our core argument is that physicists should not be mere users of AI; they should be its designers, injecting symmetry, conservation laws, causality, and interpretability into AI's DNA. This is both a technical challenge and an epistemological imperative, we should redefine what it means to understand.

The Innovation Physics champions "**See the unseen**". AI may help us see physical landscapes not yet captured by theory but only if we first enable AI to make sense, so that its outputs can be digested, questioned, and elevated by human physical intuition. This requires deep collaboration, new evaluation standards, and an unwavering commitment to physics' oldest credo: understanding goes beyond prediction.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used Deepseek to improve language and generate suggestions for Figure 1. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

1. Xu Y., Liu X., Cao X., et al. (2021). Artificial intelligence: A powerful paradigm for scientific research. *The Innovation* **2**:100179. DOI:10.1016/j.xinn.2021.100179
2. Batzner S., Musaelian A., Sun L., et al. (2022). E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **13**:2453. DOI:10.1038/s41467-022-29939-5
3. Cory-Wright R., Cornelio C., Dash S., et al. (2024). Evolving scientific discovery by unifying data and background knowledge with AI Hilbert. *Nat. Commun.* **15**:5922. DOI:10.1038/s41467-024-50074-w
4. Zeni C., Pinsler R., Zügner D., et al. (2025). A generative model for inorganic materials design. *Nature* **639**:624-632. DOI:10.1038/s41586-025-08628-5
5. Gao D., Lu S., Zhang C., et al. (2026). Autonomous closed-loop framework for reproducible perovskite solar cells.*Nature* **653**:707-714. DOI:10.1038/s41586-026-10482-y

FUNDING AND ACKNOWLEDGMENTS

This work was supported by the Beijing National Laboratory for Condensed Matter

Physics (Grant No. 2024BNLCMPKF020) and Natural Science Foundation of Hebei Province (Grant No. A2026203033). The authors would like to acknowledge Zhijun Wang, Rui Wu and Rui Zhou for insightful discussions. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

J.L., R.R. and Y.W. wrote the manuscript with figure input from Z.Z. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

The authors declare no conflicts of interest.