Interpretable Machine Learning for Predicting Functional Properties of Porous and Two-Dimensional Materials

Otis Bowman
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
otis.bowman@uab.edu

Stefano Becker
School of Computing, Clemson University, Clemson, SC, USA.
contactstefano@clemson.edu

Eduard Nane
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
eduardlane756@oregonstate.edu

Claude Webb
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
webb71@uc.edu

Abstract

The accelerating demand for advanced materials with tailored functional properties has positioned machine learning as a transformative tool in computational materials science. Porous materials, such as metal-organic frameworks and zeolites, and two-dimensional materials, including graphene and transition metal dichalcogenides, present unique predictive challenges due to their high structural diversity and complex quantum-mechanical behavior. While deep learning models have achieved remarkable accuracy in predicting properties like gas adsorption, electronic band gaps, and mechanical strength, their black-box nature raises critical concerns about scientific validity, reproducibility, and regulatory compliance. This paper develops a systems-oriented analysis of interpretable machine learning frameworks for these material classes, emphasizing the structural trade-offs between predictive performance, model transparency, computational cost, and deployment sustainability. We examine multiple interpretability architectures, from intrinsically interpretable models such as decision trees and sparse linear regressors to post-hoc explanation tools like SHAP and LIME, and evaluate their suitability for high-throughput screening and experimental validation workflows. Case illustrations drawn from recent first-principles studies on doped hexagonal boron nitride and carbon foams demonstrate how interpretability can uncover physically meaningful descriptors without sacrificing accuracy. The paper further addresses infrastructure considerations, including data standardization, model governance, and the socio-technical challenges of deploying interpretable models in both academic and industrial pipelines. Broader policy implications concerning fairness, robustness, and equitable access to material discovery technologies are discussed, and a forward-looking perspective is offered on the role of

explainable AI in accelerating the transition from computational prediction to real-world material synthesis.

Keywords

interpretable machine learning, porous materials, two-dimensional materials, functional properties, explainable AI, materials informatics, system architecture, sustainability.

1. Introduction

The discovery and optimization of materials with precisely tuned functional properties have long been a central goal of condensed matter physics and materials chemistry. Traditional approaches, rooted in trial-and-error synthesis and density functional theory calculations, are increasingly supplemented by data-driven methods that leverage large computational and experimental databases. Machine learning, in particular, has demonstrated an impressive ability to predict properties such as gas uptake capacity, electronic conductivity, and thermal stability across diverse material families [1,4,5]. However, the rapid proliferation of increasingly complex models, especially deep neural networks and ensemble methods, has introduced a fundamental tension between predictive power and interpretability. In materials science, where understanding the underlying physical mechanisms is often as important as the prediction itself, the opacity of many machine learning models undermines their acceptance by domain experts and regulatory bodies.

Porous materials and two-dimensional materials represent two of the most active frontiers in this context. Porous frameworks, ranging from zeolites to metal-organic frameworks, exhibit intricate pore geometries and surface chemistries that govern their performance in gas separation, catalysis, and energy storage. Two-dimensional materials, including graphene, hexagonal boron nitride, and transition metal dichalcogenides, display exotic electronic and optical properties that are highly sensitive to doping, strain, and layer number. Both material classes suffer from a vast chemical space that is prohibitively expensive to explore solely through computation or experiment. Machine learning offers a promising avenue for high-throughput screening, but only if the resulting predictions can be interpreted in terms of meaningful physical descriptors that guide synthesis and mechanistic understanding.

This paper adopts a systems-level perspective on interpretable machine learning for these materials, focusing not merely on algorithmic performance but on the broader architectural, infrastructural, and governance issues that determine the real-world impact of such models. We argue that interpretability is not a monolithic property but a multifaceted requirement that must be balanced against accuracy, computational cost, and deployment constraints. The following sections examine the unique predictive challenges of porous and two-dimensional materials, survey interpretability frameworks, illustrate their application with concrete examples, and discuss the trade-offs and policy implications that arise when these models are integrated into material discovery pipelines.

2. The Challenge of Predicting Functional Properties in Complex Materials

Predicting functional properties from atomic-level features is inherently difficult due to the high dimensionality of the input space and the non-linear, often quantum-mechanical, nature of the underlying relationships. In porous materials, the property landscape is characterized by a combination of topological descriptors, such as pore size and connectivity, and chemical descriptors, such as metal-ligand binding energies and accessible surface area. Two-dimensional materials present their own challenges: electronic band structures, exciton

binding energies, and catalytic activity are strongly influenced by subtle variations in lattice strain, doping concentration, and interlayer coupling. Traditional theoretical methods, while accurate, are computationally intensive and cannot scale to the millions of candidate structures that modern high-throughput synthesis can produce.

Machine learning models trained on density functional theory or experimental databases have shown that these property landscapes can be approximated with high fidelity [6,13]. Yet the models that achieve the best accuracy are often the least transparent. Deep neural networks, gradient-boosted trees, and random forests can capture complex interactions but provide little insight into which features drive a given prediction. This lack of interpretability is particularly problematic in materials science for several reasons. First, it hinders the identification of new design rules and physical heuristics that could accelerate rational material design. Second, it raises doubts about the robustness of predictions outside the training distribution, a common scenario when exploring novel chemical spaces. Third, it complicates the validation of models by experimentalists who rely on physical intuition to assess the plausibility of a computed property.

Porous materials, for instance, often exhibit counter-intuitive relationships between structural flexibility and gas selectivity, where small changes in linker geometry produce dramatic shifts in performance. Without interpretable models, it is nearly impossible to distinguish meaningful correlations from spurious ones driven by dataset biases. Similarly, in two-dimensional materials, dopant site preference and its effect on electronic structure can be obscured by the high dimensionality of the feature space. The need for interpretability thus becomes a prerequisite for scientific discovery, not merely a convenience.

3. Interpretable Machine Learning: Frameworks and Architectures

Interpretable machine learning encompasses a range of strategies that can be broadly divided into two categories: intrinsically interpretable models and post-hoc explanation techniques. Intrinsically interpretable models, such as linear regression, decision trees, and sparse additive models, offer transparency by design. Their predictions can be directly traced to the contributions of individual features, making them well-suited for applications where mechanistic understanding is paramount. However, these models often underperform on complex, non-linear property surfaces unless augmented with feature engineering or kernel transformations.

Post-hoc explanation methods, on the other hand, are applied after training a high-performance black-box model. Techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) provide local approximations to the model's decision boundary [2,3]. SHAP, grounded in cooperative game theory, allocates a contribution value to each feature for a given prediction, while LIME fits a simple surrogate model in the vicinity of the point of interest. Both methods have been widely adopted in materials informatics to reveal which structural or chemical descriptors drive predictions for specific candidate materials. Attention mechanisms in graph neural networks further enhance interpretability by highlighting which atomic neighbors or bond types are most influential [11].

Choosing the appropriate interpretability architecture depends heavily on the system context. For high-throughput screening campaigns that require rapid processing of millions of candidates, intrinsically interpretable models such as sparse linear models or small decision trees can be embedded directly into the pipeline with minimal computational overhead. In

contrast, when the goal is to understand the predictions of a state-of-the-art graph convolutional network, post-hoc methods provide a pragmatic compromise. The architectural trade-off is not merely about accuracy versus interpretability; it also involves infrastructure constraints such as memory usage, inference latency, and the ease of updating models as new data become available. Moreover, the choice of interpretability framework influences the governance and auditing capabilities of the system, a factor that becomes critical when material predictions inform safety-critical applications like carbon capture or battery performance.

4. Case Illustrations: Porous Frameworks and Two-Dimensional Layers

To ground the discussion, we consider two representative case studies from recent literature. The first involves carbon foams, a class of porous materials whose thermo-physical properties are strongly dependent on precursor raw materials. Recent work demonstrated that machine learning models trained on experimental data could predict thermal conductivity and compressive strength, but the predictive accuracy varied significantly with the choice of precursor and processing conditions [17]. Interpretability analysis using feature importance scores revealed that the carbonization temperature and the ratio of binder to filler were the dominant drivers of performance, while pore morphology played a secondary role. This insight allowed researchers to prioritize synthesis parameters that are easier to control, thereby accelerating the development of carbon foams for insulation and lightweight structural applications.

The second case concerns the gas adsorption behavior of doped hexagonal boron nitride (h-BN) monolayers, a two-dimensional material of interest for toxic gas sensing and capture. A first-principles study investigated the adsorption of CO, NH3, HCN, CNCl, and Cl2 on metal-doped and cyclic carbon-metal-doped h-BN surfaces [18]. The high computational cost of these calculations motivated the application of machine learning surrogates to predict adsorption energies across a wider space of dopant types and concentrations. Interpretability analysis via SHAP values identified that the charge transfer between the dopant atom and the gas molecule, along with the local density of states near the Fermi level, were the most consequential features. Notably, the model was able to discover a non-linear synergy between dopant electronegativity and lattice distortion that had not been explicitly encoded in the original feature set. These findings illustrate how interpretable machine learning can serve as a hypothesis-generation tool, directing experimentalists toward the most promising dopant-gas combinations for further verification.

Both cases highlight a common pattern: interpretability does not merely provide post-hoc explanations but actively shapes the experimental design and the discovery of new physical insights. In the carbon foam example, the model revealed that a relatively small set of process variables dominated the property landscape, enabling a focused experimental campaign. In the h-BN example, the model uncovered a previously overlooked interaction between electronic and geometric factors. These outcomes are only possible when the machine learning framework is designed with interpretability as a first-class citizen, not an afterthought.

5. Structural Trade-offs: Accuracy, Interpretability, and Computational Cost

The relationship between model accuracy and interpretability is often portrayed as a necessary trade-off: the most interpretable models are simple and less accurate, while the most accurate models are opaque. In the context of materials science, however, this dichotomy is moderated

by the nature of the property being predicted and the availability of domain-relevant features. For example, predicting the band gap of two-dimensional semiconductors can be achieved with high accuracy using a physically motivated descriptor such as the product of lattice constant and effective mass, resulting in a linear model that is both interpretable and performant. In contrast, predicting gas adsorption in metal-organic frameworks, where the interplay of pore shape, metal coordination, and guest-host dispersion forces is highly non-linear, may require a random forest or graph neural network to reach acceptable accuracy.

Computational cost introduces an additional dimension to this trade-off. Intrinsically interpretable models typically train faster and require less data than deep learning approaches, a critical advantage when computational resources are limited or when the training dataset is small, as is often the case in materials discovery. Post-hoc explanation methods, while enabling the use of powerful black-box models, impose an extra computational burden during inference and may themselves introduce approximation errors. In deployment scenarios, such as real-time feedback during robotic synthesis or autonomous material screening, the latency of explanation generation must be considered alongside prediction accuracy.

Furthermore, the sustainability of the machine learning pipeline—including energy consumption during training, the carbon footprint of large-scale computations, and the long-term maintainability of the model—depends on these architectural choices. Recent studies have shown that large deep learning models for materials can consume thousands of GPU hours, raising concerns about their environmental impact [9]. Interpretable models, particularly those that rely on engineered descriptors rather than learned representations, tend to be more energy-efficient. Balancing these trade-offs requires a system-level perspective that accounts for the entire lifecycle of the model, from development through deployment and eventual retirement.

6. Deployment and Infrastructure Considerations

Translating interpretable machine learning models from academic proof-of-concept to robust, production-grade material discovery platforms necessitates careful attention to infrastructure. Data standardization is a foundational challenge: property databases for porous and two-dimensional materials are heterogeneous, with varying units, measurement protocols, and levels of theoretical accuracy. Models trained on one dataset may fail when applied to another unless feature spaces are aligned and uncertainty is quantified. Reproducibility further demands that models be packaged with their training data, hyperparameter configurations, and explanation artifacts in a manner that allows independent verification.

Deployment architectures for interpretable models typically fall into two paradigms: centralized cloud-based platforms that aggregate computational and experimental data, and edge-deployable models that run on local instruments such as scanning electron microscopes or automated synthesizers. In the centralized approach, interpretability can be enhanced by providing interactive dashboards that allow scientists to query model explanations for any candidate material. In the edge paradigm, the model must be lightweight and fast, favoring intrinsically interpretable models over post-hoc explanation pipelines. The choice of deployment architecture influences not only the technical feasibility but also the governance of the model—who has access to explanations, how updates are propagated, and how model decisions are audited.

Infrastructure also includes the socio-technical systems that support model development: collaborative platforms for sharing code and data, laboratory information management

systems that feed experimental measurements back into training sets, and continuous integration pipelines that retrain models as new data arrive. Interpretability must be embedded in these workflows from the outset, with explicit documentation of which features are used, how explanations are generated, and under what conditions the model is expected to generalize. Without such infrastructure, even the most interpretable model may fail to gain trust among practitioners.

7. Sustainability, Robustness, and Fairness in Material Discovery

The long-term viability of machine learning–driven material discovery depends on three interrelated qualities: sustainability, robustness, and fairness. Sustainability encompasses the environmental cost of computation, as noted above, but also the durability of the models themselves. As databases grow and synthesis techniques evolve, models must be able to adapt without catastrophic forgetting or loss of interpretability. Robustness refers to the stability of predictions and explanations under small perturbations in input features or under shifts in the data distribution. In porous materials, for instance, a model trained on data from one class of metal-organic frameworks may perform poorly on another class because the underlying chemical mechanisms differ. Interpretability can aid robustness by highlighting when a prediction relies on features that are not transferable, enabling early detection of domain mismatch.

Fairness in material machine learning is an emerging concern. Because training datasets are often biased toward well-studied chemistries—such as transition metals with high abundance or simple pore geometries—models may systematically underperform for less-studied but potentially valuable materials. For example, two-dimensional materials composed of unconventional elements or with non-hexagonal lattices are underrepresented in computational databases, leading to biased predictions. Interpretable models allow practitioners to trace bias back to specific features, such as the presence of certain atomic species, and to correct for it through re-weighting or data augmentation. Moreover, the deployment of such models in global research communities raises questions of equitable access. If interpretable models require expensive infrastructure or proprietary data, researchers in resource-constrained settings may be excluded from participation. Open-source frameworks that emphasize interpretability and low computational cost can help democratize material discovery.

8. Policy and Governance Implications

As machine learning models become embedded in material design cycles that lead to commercial products, regulatory frameworks must evolve to address the transparency of these predictive systems. In sectors such as energy storage, carbon capture, and environmental sensing, where material performance directly impacts safety and environmental justice, the ability to audit a model’s reasoning becomes a matter of accountability. Interpretable machine learning provides the technical basis for such audits, enabling regulators and third-party evaluators to verify that predictions are grounded in physically plausible features rather than spurious correlations.

Governance models for material-focused machine learning should incorporate standards for interpretability, including minimum requirements for feature documentation, explanation fidelity, and uncertainty quantification. Professional societies, such as the Materials Research Society or the American Physical Society, could establish guidelines for the publication and peer review of machine learning–based property predictions, analogous to the reporting

standards already in place for experimental characterization. Furthermore, funding agencies and institutional review boards may need to consider the ethical implications of models that are used to prioritize materials for synthesis, particularly when those decisions affect resource allocation in large-scale research initiatives.

International collaboration is essential for developing interoperable interpretability protocols. The diverse nature of porous and two-dimensional materials research, which spans chemistry, physics, and engineering, means that no single interpretability framework will suffice. Instead, a modular governance structure that allows different communities to adopt explanations suited to their domain—while maintaining a core set of transparency principles—is likely to be most effective. These policy considerations, though often overlooked in technical papers, are critical for ensuring that interpretable machine learning fulfills its promise as a trustworthy partner in material innovation.

9. Conclusion

Interpretable machine learning offers a powerful pathway for predicting the functional properties of porous and two-dimensional materials while preserving the scientific understanding that drives discovery. This paper has examined the system-level challenges and architectural choices that shape the effectiveness of interpretability in this domain, from the selection of model class to the design of deployment infrastructure. Trade-offs between accuracy, interpretability, and computational cost must be navigated with care, and case studies from carbon foams and doped h-BN illustrate how interpretability can reveal new physical insights that guide experimental efforts. Sustainability, robustness, and fairness are integral to the long-term success of these approaches, and policy frameworks are needed to ensure that interpretability standards are upheld across the global research community. As machine learning continues to permeate materials science, the integration of interpretable architectures will be essential for translating computational predictions into tangible material advances that address pressing societal needs.

References

1. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

2. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

3. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

4. Raccuglia, P., Elbert, K. C., Adler, P. D. F., Falk, C., Wenny, M. B., Mollo, A., Zeller, M., Friedler, S. A., Schrier, J., & Norquist, A. J. (2016). Machine-learning-assisted materials discovery using failed experiments. Nature, 533(7601), 73–76. https://doi.org/10.1038/nature17439

5. Ward, L., Agrawal, A., Choudhary, A., & Wolverton, C. (2016). A general-purpose machine learning framework for predicting properties of inorganic materials. npj Computational Materials, 2, 16028. https://doi.org/10.1038/npjcompumats.2016.28

6. Xie, T., & Grossman, J. C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Physical Review Letters, 120(14), 145301. https://doi.org/10.1103/PhysRevLett.120.145301

7. Oviedo, F., Ren, Z., Sun, S., Settens, C., Liu, Z., Hartono, N. T. P., Ramasamy, S., DeCost, B., Naik, V., Mansour, M., Bousslama, S., Qi, G., Tan, H., & Buonassisi, T. (2019). Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. npj Computational Materials, 5, 60. https://doi.org/10.1038/s41524-019-0191-0

8. Agrawal, A., & Choudhary, A. (2019). Deep materials informatics: Applications and perspectives. APL Materials, 7(8), 080901. https://doi.org/10.1063/1.5111078

9. Himanen, L., Geurts, A., Foster, A. S., & Rinke, P. (2020). Data-driven materials science: Status, challenges, and perspectives. Computational Materials Science, 183, 109852. https://doi.org/10.1016/j.commatsci.2020.109852

10. Choudhary, K., DeCost, B., Chen, C., Jain, A., Tavazza, F., Cohn, R., Park, C. W., Choudhary, A., Agrawal, A., & Green, M. (2020). Recent advances and applications of deep learning methods in materials science. npj Computational Materials, 6, 157. https://doi.org/10.1038/s41524-020-00434-1

11. Chen, C., Ye, W., Zuo, Y., Zheng, C., & Ong, S. P. (2020). Graph networks as a universal machine learning framework for molecules and crystals. Journal of Chemical Information and Modeling, 60(10), 4632–4653. https://doi.org/10.1021/acs.jcim.0c00766

12. Mathew, K., Singh, A. K., Gabriel, J. J., Choudhary, K., Simnott, S. B., & Leite, M. S. (2017). A high-throughput framework for discovering new materials. Computational Materials Science, 139, 316–326. https://doi.org/10.1016/j.commatsci.2017.07.031

13. Rupp, M., Tkatchenko, A., Müller, K.-R., & von Lilienfeld, O. A. (2012). Fast and accurate modeling of molecular atomization energies with machine learning. Physical Review Letters, 108(5), 058301. https://doi.org/10.1103/PhysRevLett.108.058301

14. Faber, F. A., Lindmaa, A., von Lilienfeld, O. A., & Armiento, R. (2016). Machine learning energies of 2 million elpasolite crystals. Physical Review Letters, 117(13), 135502. https://doi.org/10.1103/PhysRevLett.117.135502

15. Gomez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., & Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. ACS Central Science, 4(2), 268–276. https://doi.org/10.1021/acscentsci.7b00572

16. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A., & Müller, K.-R. (2018). SchNet – A deep learning architecture for molecules and materials. Journal of Chemical Physics, 148(24), 241722. https://doi.org/10.1063/1.5019779

17. Al-Majali, M. R., Zhang, M., Al-Majali, Y. T., & Trembly, J. P. (2025). Impact of raw material on thermo-physical properties of carbon foam. The Canadian Journal of Chemical Engineering, 103(3), 1309-1318.

18. Zhao, J., Zhang, M., Wang, C., Yu, W., Zhu, Y., & Zhu, P. (2025). First-principles study of CO, NH3, HCN, CNCl, and Cl2 gas adsorption behaviors of metal and cyclic C–metal B-and N-site-doped h-BNs. Electronic Materials Letters, 21(2), 268-288.

19. Baird, S. G., & Sumpter, B. G. (2021). Explainable artificial intelligence for materials science. MRS Bulletin, 46, 1052–1061. https://doi.org/10.1557/s43577-021-00182-1

20. Pilania, G., Wang, C., Jiang, X., Rajasekaran, S., & Ramprasad, R. (2013). Accelerating materials property predictions using machine learning. Scientific Reports, 3, 2810. https://doi.org/10.1038/srep02810