ORIGINAL RESEARCH **Open access**

Conceptual Foundations for Adversarial Validation in Materials Machine Learning

Lucas Meyer[1*], Anna Schmid[2], Stefan Braun[1]

Abstract

Standard validation protocols in materials machine learning continue to rely on the assumption that training and test data are drawn from the same underlying distribution. This assumption is almost invariably violated in real-world materials datasets because of temporal drift in measurement techniques, compositional biases in database construction, and experimental confounders arising from different laboratories and instruments. This conceptual framework article proposes adversarial validation as a diagnostic tool specifically tailored for materials informatics: a method that trains a discriminator to explicitly detect whether a distribution shift exists between any two datasets, thereby revealing hidden generalization failures that conventional train-test splits and k-fold cross-validation cannot expose. The framework introduces the conceptual foundations of adversarial validation, distinguishes it from adversarial attacks, articulates why the technique is particularly powerful in the small-data, high-dimensional, and physically constrained domain of materials science, and offers a five-component structure for its systematic application—feature-space definition, classifier selection, shift-detection thresholding, localization of driving features, and actionable response rules. By embedding materials-specific domain knowledge into the interpretation of discriminator performance, the approach transforms validation from a passive checkpoint into an active diagnostic that can distinguish temporal shift from compositional bias and experimental confounding. The implications for materials AI practice are immediate and transformative: researchers can now report adversarial validation results alongside standard metrics, trigger targeted dataset augmentation or model retraining when shifts are detected, and document potential sources of distribution mismatch in experimental workflows, ultimately raising the robustness and trustworthiness of property predictions that underpin materials discovery and design.

Keywords Materials informatics, Distribution shift, Dataset bias, Generalization, Adversarial validation, Domain adaptation

*Correspondence:
Lucas Meyer
lucas.meyer@gmail.com
[1] Department of Materials Modeling and AI Systems, ETH Zurich, Zurich, Switzerland
[2] Department of Data-Driven Materials Engineering, University of Bern, Bern, Switzerland

Introduction

The dominant validation paradigm in materials machine learning rests on the foundational premise that training and test sets are independent and identically distributed (i.i.d.). Yet in materials science, this premise is rarely satisfied. New measurement techniques emerge, compositional coverage in databases expands unevenly, and experimental protocols differ across laboratories, all of which introduce subtle but consequential distribution shifts that standard train-test splits and cross-validation procedures cannot detect [1-5]. These blind spots become especially problematic because materials predictions are frequently deployed in high-stakes settings—battery design, catalyst screening, alloy development—where an undetected shift can lead to models that appear accurate on held-out data yet fail catastrophically when applied to new chemistries or processing conditions [6].

The present conceptual framework, therefore, reframes validation as an active diagnostic rather than a passive

quality check. It identifies adversarial validation, originally developed within broader machine-learning literature for dataset-shift detection [4], as a particularly promising instrument for materials informatics. Unlike traditional methods that only assess predictive accuracy under the i.i.d. assumption, adversarial validation explicitly tests whether any statistical difference exists between the distributions that generated the training and test (or deployment) data. When the discriminator learns to separate the two sets with accuracy significantly above chance, the framework signals that a distribution shift is present and that standard performance metrics are likely over-optimistic [3].

Figure 1 presents the proposed five-component adversarial-validation framework, illustrating how discriminator-based shift detection is embedded within a closed-loop diagnostic system linking feature definition, shift localization, and actionable response.

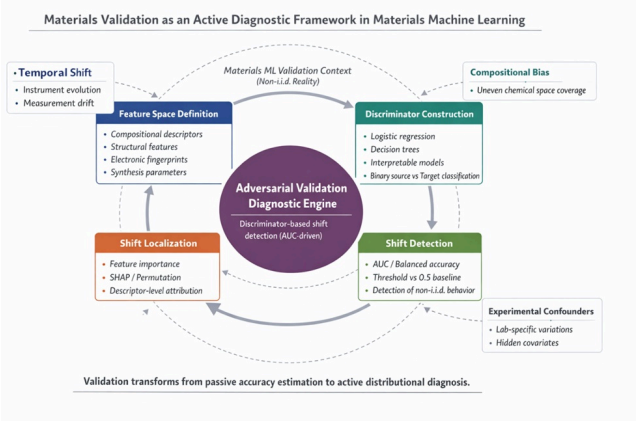

Figure 1. The proposed five-component adversarial-validation framework.

This article develops the conceptual foundations required to adapt and interpret adversarial validation for materials problems. It begins by dissecting the specific ways in which conventional validation fails in the materials domain, proceeds to a precise definition of adversarial validation itself, explains why the method aligns uniquely with the small-data and physically grounded nature of materials science, and then presents a five-component conceptual framework that researchers can apply directly. Throughout, the discussion remains strictly conceptual, focusing on logical structure, interpretive principles, and implications rather than any particular dataset or empirical benchmark. The goal is to equip the community with a new diagnostic lens that can expose hidden generalization failures and, in turn, guide more reliable model development in materials artificial intelligence.

The Validation Problem in Materials Ml

Validation practices in materials machine learning are often treated as neutral methodological safeguards. Yet, their apparent robustness conceals deeper vulnerabilities that emerge from the temporal and epistemic structure of materials data itself. One such vulnerability arises through temporal drift, where the very conditions under which data are generated evolve alongside scientific instrumentation and experimental practice. Measurement protocols are not static; they are continuously recalibrated, refined, and sometimes fundamentally redesigned. As a result, models trained on past datasets may internalize feature–property relationships that no longer hold when confronted with samples characterized using new techniques, where both measurement precision and underlying representations have shifted in subtle but consequential ways [7-14]. Under these conditions, standard cross-validation—anchored in the assumption that training and test data are drawn from the same historical distribution—fails to approximate the forward-looking deployment scenario that such models ultimately face.

This temporal instability intersects with a more structural distortion embedded in the composition of materials datasets. The distribution of data across chemical space is profoundly uneven, reflecting both historical research priorities and industrial relevance rather than any principled sampling of the periodic table or processing regimes. Well-studied systems such as oxides, perovskites, and widely used alloys dominate available datasets, while vast regions of compositional and processing space remain only sparsely explored [7]. Models trained within these dense regions can achieve impressive predictive performance when evaluated on randomly partitioned data. Yet, this apparent success is contingent upon the preservation of the same compositional biases within the validation split. When the model is extended beyond these familiar regimes, its predictive capacity often degrades sharply, revealing that what was measured as accuracy was in fact a reflection of interpolation within a highly constrained subspace rather than genuine generalization.

A further complication emerges from the aggregation of data across heterogeneous experimental environments,

where variation is not merely a matter of noise but of systematically embedded context. Differences in instrumentation, laboratory protocols, environmental conditions, and even tacit experimental practices introduce latent variables that are rarely encoded explicitly in the dataset [8-21]. These hidden confounders generate distributional discrepancies that remain invisible under i.i.d.-based evaluation schemes, yet they exert a decisive influence when models are transferred across experimental settings. In practice, the model may appear stable within the confines of its training distribution while exhibiting unpredictable behavior when confronted with data shaped by a different experimental lineage.

Taken together, these dynamics undermine the plausibility of the i.i.d. assumption as a meaningful foundation for validation in materials machine learning. What appears as methodological rigor risks becoming a ritualized confirmation of performance under artificially stabilized conditions, rather than a genuine test of model reliability in the open-ended environments of scientific discovery [18]. Addressing this limitation requires a shift in how validation itself is conceptualized: not as a passive measurement of predictive accuracy, but as an active interrogation of distributional alignment. Within this reframing, adversarial validation becomes particularly significant, as it operationalizes the detection of distributional divergence by recasting it as a classification problem, thereby exposing both the presence and structure of shifts that conventional evaluation procedures systematically overlook.

What Is Adversarial Validation?

Adversarial validation is a conceptual procedure in which a binary classifier—termed the discriminator—is deliberately trained to distinguish between two datasets that are presumed, under standard validation, to be exchangeable. One dataset is labeled “source” (typically the training set) and the other “target” (the prospective test or deployment set). If the discriminator achieves performance materially above random chance, the procedure concludes that a distribution shift exists between the two sets. The method is conceptually distinct from adversarial attacks, which seek to perturb inputs so as to induce misclassification; here, the “adversarial” label refers only to the fact that the discriminator is optimized to expose differences that a standard model would ignore [8].

Formally, let D_S and D_T denote the source and target datasets. A discriminator f_θ is trained to minimize a binary cross-entropy loss that predicts whether a given sample originates from D_S (label 0) or D_T (label 1). The validation statistic is then the area under the ROC curve (AUC) or the balanced accuracy of f_θ evaluated on a held-out portion of the combined data. An AUC near 0.5 indicates that the two distributions are empirically indistinguishable; an AUC significantly greater than 0.5 flags a detectable shift [4].

Adversarial validation for materials machine learning is the process of training a discriminator to separate training and prospective test (or deployment) data in feature space, thereby diagnosing any distribution shift that would invalidate the i.i.d. assumption underlying conventional performance metrics.

Why Adversarial Validation for Materials?

Why adversarial validation is particularly suited to materials science follows from three structural characteristics of the domain. Materials datasets are characteristically small, often containing only hundreds or low thousands of labeled examples for many property-prediction tasks. Under such conditions, classical statistical tests for distribution equality lack statistical power [19]. Adversarial validation circumvents this limitation by recasting shift detection as a supervised classification problem, remaining statistically efficient even in modest-sample regimes. The discriminator exploits the full feature space to uncover subtle distributional differences—differences that a univariate Kolmogorov–Smirnov test would miss.

A related implication stems from the high-dimensional and physically structured nature of materials feature spaces. Compositions, crystal descriptors, electronic-structure fingerprints, and synthesis parameters all coexist within the same input vector, and this dimensionality generates a combinatorial explosion of possible shift directions [12]. Conventional validation splits offer no systematic way to enumerate or localize these directions. The adversarial discriminator, in contrast, naturally surfaces the most discriminative axes, thereby converting what would otherwise remain an opaque generalization failure into an interpretable, feature-level diagnosis.

Beyond this immediate diagnostic utility, materials science provides strong domain priors—physical laws, periodic trends, thermodynamic constraints—that can be injected directly into the interpretation step. When the discriminator flags a shift, the materials expert can map the driving features onto known physical mechanisms: a change in oxidation-state distribution, a shift in synthesis-temperature histograms, or similar phenomena. Raw statistical detection thus becomes actionable scientific insight [15]. Under these conditions, what might serve as a generic machine-learning diagnostic transforms into a materials-specific instrument—one capable of revealing precisely those generalization failures that matter most for trustworthy property prediction.

A Conceptual Framework for Adversarial Validation

The conceptual framework for adversarial validation in materials machine learning rests on several interlocking components that together convert raw distributional difference detection into a structured, interpretable, and actionable diagnostic. A logical starting point is the explicit declaration of the descriptor set on which the discriminator will operate. Because different descriptor families capture fundamentally different physical mechanisms—compositional versus structural versus electronic, for instance—the choice of feature space directly determines which classes of shift the procedure can detect [16]. The framework, therefore, mandates that researchers document the exact feature representation used for validation and, when practicable, run parallel validations across complementary descriptor families to triangulate the nature of any detected shift.

Once the feature space is fixed, the choice of validation classifier becomes equally consequential. In materials contexts, interpretability is paramount; accordingly, the framework recommends simple, intrinsically interpretable models—logistic regression, shallow decision trees, or small random forests—rather than complex neural discriminators [13]. Simplicity ensures that feature importances and decision paths remain directly mappable to physical quantities, allowing domain experts to translate statistical findings into materials-specific hypotheses without resorting to additional post-hoc explanation techniques. This preference for transparency also informs how the framework treats the discriminator's performance. Rather than imposing a binary pass/fail threshold, the framework conceptualizes the AUC or balanced accuracy as a continuous diagnostic statistic. A conceptual baseline is set by reference to the true separability expected under genuine i.i.d. sampling, approximately 0.5, adjusted upward only when domain knowledge suggests that minor, physically inconsequential differences are tolerable [20-23]. The precise numerical cutoff remains a tunable hyperparameter whose justification must be reported alongside any validation result.

A related implication concerns what happens after a shift is detected. The framework requires systematic localization: permutation importance, SHAP values, or partial-dependence profiles are extracted from the discriminator to identify which specific descriptors drive the separation [24-26]. This step transforms a global detection signal into a feature-level map, one that can be aligned with known materials mechanisms such as temporal instrument drift, compositional undersampling, or laboratory-specific confounders. Localization, in turn, feeds directly into the final component: codified decision rules that link detected shifts to concrete modeling actions. Under this scheme, if no shift is detected, standard validation proceeds; if a shift is localized to a particular feature family, targeted data augmentation or reweighting is triggered; and if the shift is global and severe, model retraining on an expanded or rebalanced dataset is recommended [20]. These rules are framed conceptually so that practitioners can adapt them to the risk tolerance and downstream application of any given materials prediction task. Taken together, these components—feature-space definition, classifier choice, performance thresholding, shift localization, and action rules—provide a complete conceptual scaffold, elevating adversarial validation from an ad-hoc diagnostic into a repeatable, transparent, and materials-aware validation protocol.

Interpreting Adversarial Validation Results

Once a discriminator signals the presence of a distribution shift, the interpretive phase transforms the raw separability statistic into materials-specific insight by examining the pattern of discrimination rather than merely its existence.

To operationalize the interpretation of discriminator outcomes, **Table 1** maps characteristic adversarial-validation patterns to underlying materials-specific failure modes and corresponding corrective actions.

Table 1. Diagnostic mapping of adversarial validation outcomes to materials-specific failure modes and corrective actions

Discriminator pattern	AUC behavior	Feature attribution pattern	Interpreted shift type
Near-random separability	~0.5	No dominant features	No detectable shift
Uniform high separability	High across all features	Diffuse importance	Global distribution shift
Feature-specific separability	Moderate–high	Concentrated on a subset	Localized shift
Temporal gradient in AUC	Increasing with a time gap	Time-correlated features	Progressive temporal shift
Weak but significant separability	Slightly > 0.5	Low-intensity signals	Minor shift (possibly noise)
No separability, but poor performance	~0.5	No signal	Conditional shift (label relationship)

Three archetypal patterns emerge that the conceptual framework distinguishes and maps onto known failure modes in materials informatics. Uniform separability occurs when the discriminator achieves high performance across the entire feature space without any single descriptor dominating; this pattern conceptually identifies a global, systematic shift such as a wholesale change in measurement protocols or a broad temporal drift that affects every dimension of the data [14]. In contrast, feature-specific separability arises when permutation importance or SHAP-derived attributions concentrate on a narrow subset of descriptors—perhaps only the compositional fingerprints or only the electronic-structure features—thereby localizing the shift to a particular physical mechanism such as undersampling of certain oxidation states or a bias toward particular crystal symmetries [26].

A third pattern, temporal trend in separability, appears when the discriminator is retrained successively on data binned by publication or synthesis year; an increasing AUC as the temporal gap widens reveals a progressive divergence that standard cross-validation, which mixes all years indiscriminately, would never expose [17]. Materials domain knowledge plays a decisive role in translating these statistical patterns into scientific understanding. For instance, if the driving features identified by the discriminator align with known changes in X-ray diffraction resolution across instrument generations, the researcher can confidently attribute the shift to instrumental evolution rather than to any intrinsic materials phenomenon [21].

The framework further proposes that interpretation should proceed through a structured diagnostic loop: first, quantify the global AUC, then extract localized feature attributions, and finally overlay physical priors—thermodynamic stability windows, periodic-table trends, or synthesis-parameter ranges—to decide whether the detected shift is practically consequential [23]. When uniform separability coincides with a large temporal gap, the interpretation suggests that the entire historical database may require reweighting or augmentation before further model deployment. When feature-specific separability points to compositional bias, the framework recommends targeted acquisition of underrepresented chemistries rather than blanket data collection. This interpretive layer thus elevates adversarial validation from a mere alarm system to a diagnostic that guides precise corrective actions, ensuring that detected shifts inform both model refinement and experimental planning without ever claiming empirical performance on any particular dataset.

Relationship to Existing Validation Methods

Adversarial validation occupies a distinct conceptual niche among validation strategies in materials machine learning and complements rather than supplants established techniques. Cross-validation and random train-test splits operate entirely within the i.i.d. assumption, measuring predictive accuracy while remaining blind to the very distributional differences that adversarial validation is designed to expose [18]; they therefore remain essential for estimating in-distribution performance yet offer no diagnostic power when that assumption fails. Out-of-distribution detection methods, by contrast, typically evaluate model confidence or reconstruction error on held-out points. Still, they assess the model's reaction to a shift rather than testing whether a shift exists in the data-

generating process itself [27-29]. Domain adaptation techniques presuppose that a shift has already been identified and then attempt to mitigate it through reweighting or feature alignment; adversarial validation precedes and informs such adaptation by providing an explicit, quantitative confirmation that adaptation is warranted [2]. To clarify the distinct diagnostic role of adversarial validation relative to existing validation paradigms, **Table 2** provides a structured conceptual comparison across key methodological dimensions

Table 2. Conceptual comparison of validation paradigms in materials machine learning

Dimension	Cross-validation (i.i.d.)	Out-of-distribution detection	Domain adaptation
Core objective	Estimate predictive accuracy	Detect model uncertainty under shift	Mitigate known shift
Assumption	Training = test distribution	Shift may exist post hoc	Shift already identified
What is tested	Model performance	Model response to anomalies	Model adaptation effectiveness
Sensitivity to small data	Moderate	Limited	Variable
Interpretability	Low (scalar metrics)	Low–moderate	Low
Localization of shift	None	None	Partial
Actionability	Minimal	Limited	Reactive
Role in pipeline	Standard baseline	Downstream check	Corrective step

The conceptual comparison can be articulated along five dimensions. First, regarding what is being tested: adversarial validation directly probes distributional equality between source and target data, whereas cross-validation tests predictive fidelity under an assumed equality. Second, regarding sensitivity to small-data regimes: adversarial validation remains statistically efficient because it reframes the problem as supervised classification, while classical statistical tests for distribution equality lose power rapidly as sample size decreases [19]. Third, regarding interpretability: the discriminator's feature attributions yield localized insights into which descriptors drive the mismatch, an advantage not shared by scalar performance metrics from cross-validation or global uncertainty scores from out-of-distribution detectors. Fourth, regarding actionability: adversarial validation supplies explicit decision rules that trigger dataset curation or model retraining, whereas domain-adaptation pipelines require prior knowledge that a shift exists. Fifth, regarding integration with domain knowledge: adversarial validation invites the injection of materials-specific priors during interpretation, turning statistical detection into a physically grounded diagnosis in a way that purely model-centric methods cannot.

Thus, the framework positions adversarial validation as a necessary upstream diagnostic that reveals the limitations of existing methods and supplies the missing precondition for their responsible application. When used in concert with cross-validation, out-of-distribution testing, and domain adaptation, it forms a layered validation strategy that addresses both in-distribution accuracy and out-of-distribution robustness without redundancy [4].

Limitations and Open Questions

Despite its power as a conceptual lens for diagnosing distribution shift, adversarial validation cannot detect every form of generalization failure that may arise in materials machine learning. The method operates strictly on the observed feature space; consequently, any shift that manifests only in the label distribution or in the conditional relationship between features and properties—while the

marginal feature distribution remains indistinguishable—will remain invisible to the discriminator [3]. A related vulnerability concerns false positives: minor, physically inconsequential differences, such as slight variations in measurement noise floors, can produce statistically detectable separability that does not materially affect downstream property prediction. False negatives, conversely, may emerge when the chosen feature space or discriminator architecture lacks sufficient capacity to capture subtle but consequential shifts.

These conceptual limitations point toward several avenues for future refinement. Principled thresholds for actionable intervention remain to be formalized: how large must the discriminator's AUC be, and how localized must the driving features become, before model retraining or dataset expansion is mandated? Under current practice, these decisions remain largely heuristic. A related opportunity lies at the intersection of active learning protocols. Adversarial validation could, in principle, guide the selection of new experiments that maximally reduce detected shifts rather than merely maximizing model uncertainty [15]. That shift from uncertainty-driven to shift-driven experimental design remains largely unexplored. Beyond this, real-world materials data often arrive from more than two distinct experimental campaigns, introducing multi-source shift. Extending the framework to handle such settings will require a generalized approach capable of disentangling simultaneous temporal, compositional, and laboratory-specific contributions.

The present conceptual framework, therefore, identifies these boundaries explicitly, not as weaknesses to be concealed but as conditions requiring epistemic humility. Practitioners are best served by deploying adversarial validation as one indispensable instrument within a broader validation ecosystem—a powerful but inherently bounded tool rather than a universal solution.

Implications for Materials AI Practice

Adopting the adversarial-validation framework would fundamentally alter how the materials community conducts and reports machine-learning research. Researchers should report discriminator AUC and feature attributions alongside conventional metrics such as mean absolute error on random splits, thereby providing readers with a transparent assessment of distributional robustness rather than an incomplete picture of in-distribution accuracy [6]. When shifts are detected, the framework recommends explicit documentation of the suspected physical origins—temporal instrument evolution, compositional undersampling, or laboratory confounders—so that subsequent studies can address the same sources systematically.

In experimental workflows, detected shifts should trigger targeted actions: augmentation of underrepresented chemistries when compositional bias is localized, recalibration or harmonization of legacy data when temporal trends appear, or stratified sampling across laboratories when experimental confounders dominate [20]. Model-development pipelines would incorporate adversarial validation as a routine gate before deployment, ensuring that property predictions intended for high-stakes applications carry an accompanying statement of distributional reliability.

Collectively, these practice-level changes elevate validation from a post-hoc ritual to an integral part of the scientific method in materials informatics, fostering greater reproducibility, reduced over-optimism, and more efficient allocation of experimental resources toward closing the very gaps that adversarial validation reveals.

Conclusion

This conceptual framework has articulated the foundations of adversarial validation as a diagnostic tool uniquely suited to the small-data, high-dimensional, and physically grounded challenges of materials machine learning. By training a discriminator to expose distribution shifts that conventional i.i.d.-based validation cannot detect, the approach distinguishes temporal drift, compositional bias, and experimental confounding with a precision and interpretability that existing methods lack. The five-component structure—feature-space definition, classifier choice, performance thresholding, shift localization, and action rules—provides a repeatable scaffold that transforms raw detection into actionable scientific insight. When integrated with domain knowledge and used alongside cross-validation and out-of-distribution testing, adversarial validation supplies the missing diagnostic layer that materials AI practice has long required.

The framework therefore calls for adversarial validation to become standard reporting practice in every materials

machine-learning study, ensuring that published models carry not only accuracy metrics but also explicit evidence of distributional robustness. Only through such rigorous, conceptually grounded validation can the community build the trustworthy, generalizable models that will accelerate materials discovery and design for the decades ahead.

Acknowledgements

None

Conflict of interest

None

Financial Support

None

Ethics statement

None

Received: 20 Jul 2021 Revised: 31 Aug 2021 Accepted: 18 Oct 2021
Published online: 18 January 2022

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun ACM. 2020;63(11):139-44.

Niu S, Liu Y, Wang J, Song H. A decade survey of transfer learning (2010-2020). IEEE Trans Artif Intell. 2021;1(2):151-66.

Nasteski V. An overview of the supervised machine learning methods. Horizons. 2017;4:51-62.

Kulinski S, Bagchi S, Inouye DI. Feature shift detection: Localizing which features have shifted via conditional distribution tests. NeurIPS. 2020;33:19523-33.

Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018;559(7715):547-55.

Schmidt J, Marques MR, Botti S, Marques MA. Recent advances and applications of machine learning in solid-state materials science. npj Comput Mater. 2019;5(1):83.

Cheng Y, Zhu L, Wang G, Zhou J, Elliott SR, Sun Z. Vacancy formation energy and its connection with bonding environment in solid: A high-throughput calculation and machine learning study. Comput Mater Sci. 2020;183:109803.

Biggio B, Roli F. Wild patterns: Ten years after the rise of adversarial machine learning. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security; 2018. p. 2154-6.

Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP); 2017. p. 39-57.

Eykholt K, Evtimov I, Fernandes E, Li B, Rahmati A, Xiao C, et al. Robust physical-world attacks on deep learning visual classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 1625-34.

Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A. Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security; 2017. p. 506-19.

Garg V, Jegelka S, Jaakkola T. Generalization and representational limits of graph neural networks. In: International Conference on Machine Learning; 2020. p. 3419-30.

Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, et al. Graph neural networks: A review of methods and applications. OAI. 2020;1:57-81.

Fujinuma N, DeCost B, Hattrick-Simpers J, Lofland SE. Why big data and compute are not necessarily the path to big materials science. Commun Mater. 2022;3(1):59.

Wang AY, Mahmoud MS, Czasny M, Gurlo A. CrabNet for explainable deep learning in materials science: Bridging the gap between academia and industry. Integr Mater Manuf Innov. 2022;11(1):41-56.

Sun B, Wu Z, Hu Y, Li T. Golden subject is everyone: A subject transfer neural network for motor imagery-based brain computer interfaces. Neural Netw. 2022;151:111-20.

Koh PW, Sagawa S, Marklund H, Xie SM, Zhang M, Balsubramani A, et al. Wilds: A benchmark of in-the-wild distribution shifts. In: International Conference on Machine Learning; 2021. p. 5637-64.

Chibani S, Coudert FX. Machine learning approaches for the prediction of materials properties. APL Mater. 2020;8(8).

Mahanti R. Data quality: Dimensions, measurement, strategy, management, and governance. QP. 2019.

Habsah F. Developing teaching material based on realistic mathematics and oriented to the mathematical reasoning and mathematical communication. JRPM. 2017;4(1):43-55.

Ghiringhelli LM, Rossi M. Reliable quantification of uncertainties: The biggest challenge for data-centric materials modeling? RCMS. 14.

Gigli L, Veit M, Kotiuga M, Pizzi G, Marzari N, Ceriotti M. Thermodynamics and dielectric response of BaTiO3 by data-driven modeling. npj Comput Mater. 2022;8(1):209.

Steinhardt J. Robust learning: Information theory and algorithms. Stanford University; 2018.

Xie J, Su Y, Xue D, Jiang X, Fu H, Huang H. Machine learning for materials research and development. Acta Metall Sin. 2021;57(11):1343-61.

Chuah J, Wang M. Framework for testing robustness of machine learning-based classifiers. Sensors. 2022;22(17):1-20. https://doi.org/10.3390/s22176400.

Doshi-Velez F, Kim B. Considerations for evaluation and generalization in interpretable machine learning. In: Explainable and interpretable models in computer vision and machine learning. Cham: Springer; 2018. p. 3-17.

Cui P, Zhang Y. Out-of-distribution (OOD) detection based on deep learning: A review. Electronics. 2022;11(21):3500. https://doi.org/10.3390/electronics11213500.

Teso S, Kersting K. Explanatory interactive machine learning. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. ACM; 2019. p. 239-45.

Lehmann ML, Yang G, Gilmer D, Han KS, Self EC, Ruther RE, et al. Tailored crosslinking of poly(ethylene oxide) enables mechanical robustness and improved sodium-ion conductivity. Energy Storage Mater. 2019;21:85-96.