ORIGINAL RESEARCH Open access

A Conceptual Theory of Measurement Validity for AI-Generated Materials Properties

Andreas Müller[1*], Stefan Weber[1], Julia Hoffmann[2], Lukas Schneider[1], Tobias Klein[2]

Abstract

The pervasive reliance on predictive accuracy metrics such as mean absolute error, root mean square error, and R^2 in materials artificial intelligence has created a fundamental misconception: that low prediction error equates to a valid measurement of a material's property. This paper argues that accuracy alone is insufficient because an AI-generated property value may align closely with held-out test data yet fail to support the specific scientific or engineering inferences for which it is intended. Drawing on foundational measurement validity theory from psychometrics and the social sciences, the manuscript adapts these concepts to the unique context of AI-generated materials properties. It proposes a novel five-component conceptual theory of measurement validity tailored to machine-learning predictions of physical quantities such as band gaps, formation energies, and mechanical moduli. Five distinct dimensions of validity—construct, criterion, generalizability, robustness, and consequential—are articulated and illustrated with materials-specific scenarios. Finally, the framework offers concrete implications for authors, reviewers, and the broader materials informatics community, shifting validation practices from narrow accuracy reporting toward comprehensive evidence-based arguments that link predictions to intended uses. By distinguishing accuracy from validity, this conceptual framework aims to elevate the epistemological rigor of AI-driven materials discovery and design.

Keywords Materials informatics, Validation framework, Measurement validity, AI-generated materials properties, Construct validity, Accuracy versus validity

*Correspondence:
Andreas Müller
andreas.mueller@gmail.com
[1] Department of Materials Informatics and AI Systems, Heidelberg University, Heidelberg, Germany
[2] Department of Computational Materials Engineering, Technical University of Munich, Munich, Germany

Introduction

Artificial intelligence has transformed the discovery and design of new compounds by promising rapid, high-throughput prediction of properties that would otherwise require costly and time-consuming experiments. Across thousands of published studies, the dominant validation strategy remains the reporting of accuracy metrics—mean absolute error, root mean square error, coefficient of determination—computed on held-out test sets or through cross-validation [1-5]. Yet this narrow focus obscures a deeper epistemological issue: accuracy is not the same as validity.

Table 1 systematically distinguishes accuracy-based evaluation from measurement validity, clarifying why predictive performance alone cannot justify the interpretation or use of AI-generated materials properties."

Table 1. Distinguishing accuracy from measurement validity in AI-generated materials properties

Dimension	Accuracy-centric evaluation	Measurement validity framework	Conceptual contribution
Core definition	Numerical closeness to reference data	Evidence-based justification of interpretation and use	Reframes validation as epistemic judgment
Unit of assessment	Prediction error (MAE, RMSE, and R^2)	Validity argument (theory + evidence integration)	Moves beyond metric-based evaluation
Construct representation	Implicit/assumed	Explicitly theorized and justified	Prevents proxy-based mismeasurement
Relation to use	Typically unspecified	Central (use argument defines requirements)	Aligns model evaluation with the decision context
Evidence type	Single-metric dominance	Multi-source (accuracy, robustness, theory, and experiments)	Introduces evidentiary pluralism
Generalizability	Limited to test distribution	Explicitly evaluated across domains and conditions	Addresses out-of-distribution validity
Robustness	Error stability under perturbation	Stability of interpretive meaning under perturbation	Shifts focus from performance to meaning
Consequences	Ignored	Explicitly assessed	Introduces ethical and practical accountability
Temporal nature	One-time validation step	Ongoing evaluative process	Establishes dynamic validation logic
Epistemological status	Statistical performance claim	Scientific measurement claim	Elevates the rigor of AI outputs

A prediction may be numerically close to a reference value and still be invalid if it misrepresents the underlying physical construct, fails under conditions relevant to the intended application, or produces unintended scientific or societal consequences.

The problem is not merely technical but conceptual. When an AI model predicts the bandgap of a perovskite with a mean absolute error of 0.1 eV on a benchmark dataset, researchers routinely conclude that the model is “validated.” Such claims assume that statistical fidelity to a test distribution automatically licenses the use of those predictions for downstream tasks such as screening candidates for photovoltaics or informing synthesis decisions. This assumption is rarely examined. The present work, therefore, adapts measurement validity theory—originally developed in psychometrics and the philosophy of social science—to the domain of AI-generated materials properties [1, 2].

Validity, in this adapted sense, concerns the degree to which evidence and theoretical reasoning support the interpretation and use of an AI-generated property value for a proposed purpose. The manuscript proceeds in five stages. First, it traces the historical and conceptual origins of measurement validity. Second, it surveys current validation practices in materials artificial intelligence and demonstrates their exclusive emphasis on accuracy. Third, it articulates why accuracy is conceptually and practically insufficient. Fourth, it proposes a five-component theory of measurement validity specifically for AI-generated properties. Fifth, it lays the groundwork for subsequent sections by outlining how the theory will be elaborated through dimensions, relations to existing concepts, and practical implications. Throughout, the argument maintains a strictly conceptual orientation, avoiding any empirical claims, datasets, or performance numbers. Instead, it offers a framework for reframing validation as an ongoing, purpose-driven scientific inquiry rather than a one-time accuracy check. This shift is essential if materials artificial intelligence is to move beyond correlative prediction toward trustworthy, interpretable, and actionable knowledge [3].

Measurement Validity: Origins and Concepts

Measurement validity theory emerged within psychometrics and the social sciences as a response to the realization that a test score could be statistically reliable yet still fail to capture the intended theoretical construct. The canonical modern definition [1] states that validity is "the degree to which evidence and theory support the interpretation of test scores for proposed uses." This definition moved validity away from a property of the test itself and toward an evaluative judgment about the inferences and decisions based on the scores. Construct validity was earlier formalized [2] as the central concern: whether the measurement truly reflects the underlying theoretical entity it purports to represent.

Four interrelated types of validity evidence have become standard in the literature. Construct validity asks whether the instrument adequately embodies the theoretical construct. Criterion validity examines correlations with external, independently measured outcomes that the construct should predict. Content validity evaluates whether the full domain of the construct is represented without omission or irrelevance. Consequential validity considers the broader societal, ethical, or scientific consequences of using the measurement in decision-making. These categories are not mutually exclusive; they form an integrated argument that must be assembled case by case [1].

When transposed to artificial intelligence for materials science, the same logic applies with heightened urgency. An AI-generated property—such as a predicted elastic modulus or thermal conductivity—is not a direct physical observation but a model-derived inference. The numerical output, therefore, functions analogously to a test score: its meaning and legitimacy depend on the interpretive argument that links the model's internal representations to the target physical quantity and on the evidence that this link holds for the intended use. Validity theory thus supplies a principled language for asking whether an AI prediction of, say, the formation energy of a high-entropy alloy truly captures the thermodynamic stability construct or merely correlates with it through dataset-specific artifacts [3].

The theory also emphasizes that validity is never absolute. It is always relative to a specific context of use and must be supported by multiple lines of evidence rather than a single metric. This perspective directly challenges the prevailing assumption in materials artificial intelligence that a low error on a test set constitutes sufficient validation. By drawing explicitly on the framework in [1, 2], the present conceptual theory offers materials scientists a vocabulary and logic that have been absent from most machine-learning studies in the field.

Current Validation Practices in Materials AI

Contemporary literature in materials artificial intelligence overwhelmingly equates validation with the demonstration of predictive accuracy on held-out data. One prominent study [4] exemplifies this approach by benchmarking machine-learning models against experimental or high-fidelity computational results and declaring success when error metrics fall within acceptable ranges. Similarly, another major review [5] repeatedly highlights cross-validation scores and out-of-sample R^2 values as the primary indicators of model quality. While acknowledging pitfalls, a further analysis [6, 7] still frames validation primarily through accuracy on test sets drawn from the same distribution as the training data.

This pattern repeats across dozens of studies. A critical examination [8] evaluates robustness and generalizability by examining error metrics under modest distribution shifts, yet the ultimate judgment remains whether the mean absolute error remains low. Studies that apply machine learning to material synthesis and property prediction [9] validate models solely through agreement with reference databases. A comprehensive survey [10] cites test-set performance as the decisive validation criterion. The introduction of a RESTful API for property predictions [11] validates the service by comparing outputs to known benchmarks using standard regression statistics. Research that asks whether machine learning can identify extraordinary materials [12] answers the question with accuracy-based screening results.

Additional examples reinforce the uniformity [13-17]. Work on explainable AI for material property prediction [15] grounds its claims entirely in predictive-error reductions. The application of domain adaptation [16] reports improved R^2 and MAE as proof of validity. The introduction of dataset redundancy control [17] demonstrates its benefit through lowered test-set

errors. Across these and many parallel works [4, 5, 7-12, 15-17], the validation pipeline is identical: split data, train, compute accuracy metrics on the hold-out portion, and conclude that the model is validated for the property in question.

What is conspicuously absent is any discussion of construct representation, appropriateness for downstream decisions, or potential consequences. No study in the surveyed corpus articulates an interpretive argument about what the AI-generated value actually means physically. None systematically evaluates whether the prediction supports the specific use case—screening, inverse design, or synthesis guidance—for which it is generated. Experimental follow-up is mentioned only sporadically and never as required validity evidence. The result is a literature in which thousands of models are declared “validated” without ever addressing whether the predictions are measuring the right thing in the right way for the right purpose. This narrow accuracy-centric paradigm is precisely the gap that measurement validity theory is designed to fill [1, 3].

The Problem: Accuracy ≠ Validity

Accuracy and validity are often treated as interchangeable within materials artificial intelligence, yet this equivalence obscures a fundamental epistemic distinction that becomes especially consequential in scientific discovery contexts. High predictive performance can coexist with a misalignment between what a model captures and the physical construct it is presumed to represent. A bandgap predictor, for instance, may achieve low error across a large dataset of inorganic compounds while relying on latent correlations tied to atomic number or averaged electronegativity, rather than encoding the underlying quantum-mechanical structure that defines the property itself. Under such conditions, numerical agreement within the training distribution masks a deeper representational failure, one that becomes visible only when the model is extended into unfamiliar chemical regimes where the learned proxy relationships no longer hold [4, 5].

This tension becomes more pronounced when predictive success is interpreted as evidence of practical utility. A model calibrated to formation energies with sub-50 meV/atom error may perform convincingly within a narrowly defined compositional family. Yet, its outputs can become misleading when extrapolated to synthesis contexts governed by non-equilibrium dynamics. The apparent fidelity of the predictions invites their use in experimental decision-making, even though the conditions under which the model was trained do not align with those under which it is applied. What appears as technical precision therefore risks becoming a source of epistemic overreach, particularly when the boundary between interpolation and extrapolation is not explicitly acknowledged.

Beyond questions of representation and applicability, a further complication emerges from the downstream consequences of model deployment. Predictive accuracy, as conventionally measured, remains silent on the effects of acting on those predictions. A model that systematically underestimates fracture toughness in high-entropy alloys may still register strong aggregate performance. Yet, its outputs can encourage material selections that are unsafe or suboptimal in practice. The absence of explicit attention to uncertainty, model limitations, and context-specific risk transforms accurate predictions into potentially hazardous recommendations, revealing a gap between statistical performance and decision-relevant reliability [8].

Taken together, these considerations indicate that accuracy functions only as a preliminary condition for assessing model quality rather than a definitive criterion of validity. In materials AI, where predictions are frequently treated as surrogates for measurement, this distinction carries particular weight. Proxy-driven estimates of thermal conductivity may degrade sharply outside familiar chemistries, screening tools that perform well under standard conditions may fail under industrially relevant pressures or temperatures, and generative models can produce outputs that appear chemically plausible while violating thermodynamic constraints when scrutinized more closely. Such cases underscore the need to reconceptualize validation as an ongoing epistemic practice rather than a static performance check. A more robust framework would treat AI-generated properties as claims that require justification through alignment with physical theory, sensitivity to context, and continuous empirical interrogation, ensuring that their scientific legitimacy is argued and sustained rather than inferred from accuracy alone [1, 3].

A Theory of Measurement Validity for AI-Generated Properties

This study proposes a conceptual theory of measurement validity specifically adapted to AI-generated materials properties. The theory rests on five interlocking components that together form a coherent validity argument [3].

Figure 1 presents a hierarchical conceptualization of the shift from accuracy-centric validation toward a structured measurement validity framework grounded in interpretive and use-based reasoning.

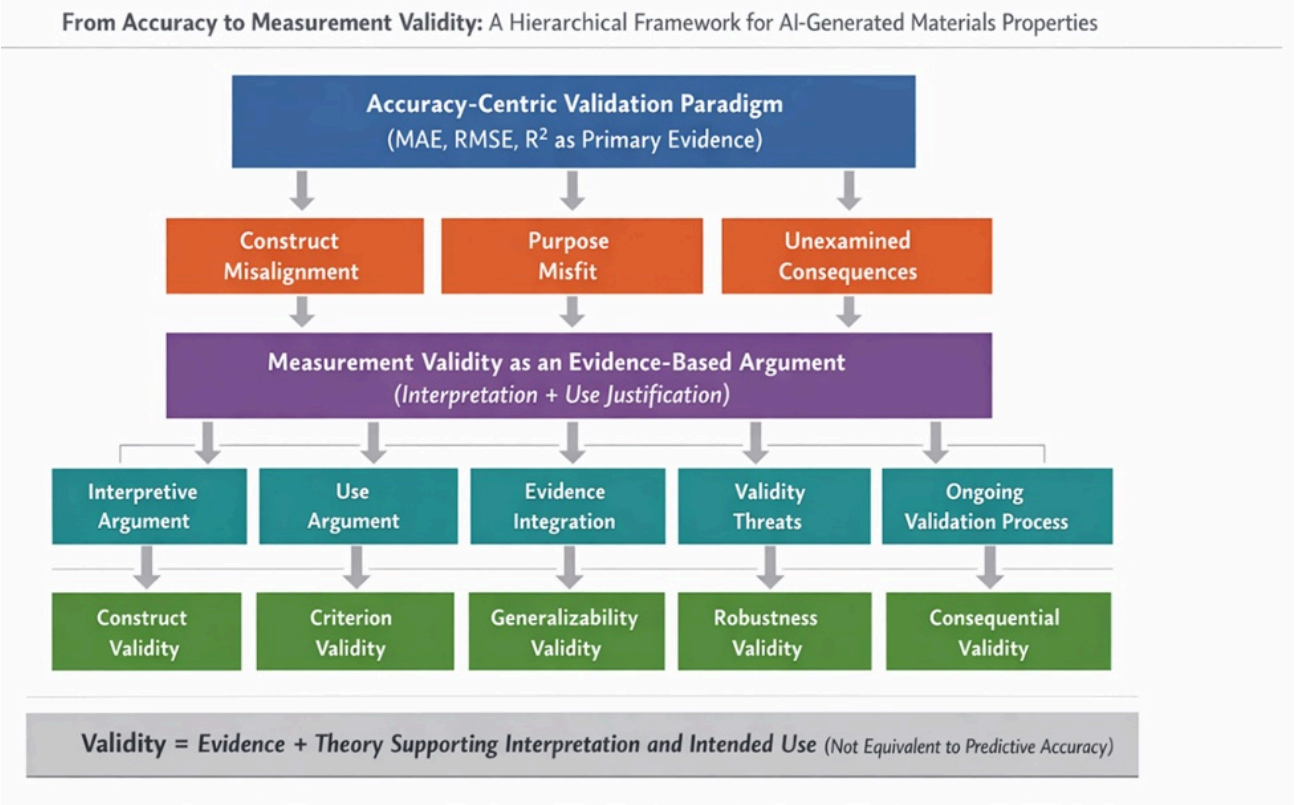

Figure 1. From accuracy to measurement validity

Any AI-generated materials property implicitly advances an interpretive claim about the physical quantity it purports to represent, yet this claim often remains unarticulated in practice. A defensible validity framework requires that the meaning of the output be made explicit and theoretically grounded, not merely assumed from training objectives or dataset labels. When a neural network produces a value described as formation energy, for example, the argument must clarify the thermodynamic conditions, computational reference (such as DFT at 0 K under standard pressure), and the extent to which the learned representation captures that construct rather than a statistically convenient proxy. The critical issue is not whether the model predicts well, but whether its internal regularities can plausibly be interpreted as encoding the intended physical phenomenon.

This interpretive commitment immediately anchors the question of use, since validity is inseparable from the purposes for which predictions are mobilized. The same output may be acceptable for coarse-grained screening yet inadequate for guiding synthesis or informing multiscale simulations where error tolerances and failure costs differ substantially. Establishing validity, therefore, requires an explicit articulation of the decisions that will depend on the model and the performance thresholds necessary to support those decisions. Under these conditions, predictive outputs become situated within a broader inferential context, where their adequacy is judged relative to the demands of specific scientific or engineering tasks rather than abstract accuracy metrics.

A credible validity claim cannot rest on a single evidentiary dimension, as no isolated metric is sufficient to establish that a model both represents a construct faithfully and supports its intended use. Instead, multiple lines of evidence must be brought into alignment, combining statistical performance with demonstrations of robustness under distributional variation, coherence with established physical principles, and, where available, correspondence with experimental observations. The crucial requirement is not the presence of these elements in isolation, but their integration into a coherent argument that links representation, performance, and application. Without such integration, validation remains fragmented and unable to support strong inferential claims.

At the same time, any validity argument must contend with the conditions under which it may fail. Materials AI systems are particularly susceptible to dataset imbalances, architectural biases, and extrapolative errors that are not readily visible through conventional evaluation protocols. A rigorous approach, therefore, requires the systematic identification of potential threats that could undermine either the interpretive grounding or the intended use of the model. Recognizing these vulnerabilities does not weaken the validity claim; rather, it situates it within realistic epistemic boundaries and opens the possibility of mitigation strategies or bounded deployment.

Even when supported by carefully constructed arguments and converging evidence, validity cannot be treated as a fixed attribute of a model. As new data become available, as applications shift, and as theoretical understanding evolves, the conditions that justified earlier interpretations may no longer hold. Validation must therefore be understood as an ongoing scientific practice in which claims are continually reassessed and refined. This perspective repositions evaluation from a terminal checkpoint to a dynamic process that evolves alongside the systems it seeks to justify.

Taken together, these elements form a unified conceptual structure in which interpretive grounding and intended use provide the foundational orientation, evidentiary integration and threat recognition stabilize the argument, and continuous re-evaluation ensures its adaptability over time. In this light, measurement validity for an AI-generated materials property can be understood as the extent to which accumulated evidence and theoretical reasoning support interpreting the model output as a representation of a specified physical construct while justifying its deployment for a defined scientific or engineering purpose [1].

Dimensions of Validity for Materials AI

The proposed theory reframes measurement validity in materials artificial intelligence as a structured yet inherently integrative judgment, grounded in a set of interdependent dimensions that together render model outputs scientifically interpretable and practically defensible. What distinguishes this formulation from conventional evaluation schemes is its explicit recognition that AI-generated properties function as inferential surrogates rather than direct measurements, emerging from learned representations of chemical and structural data rather than physical instrumentation. This shift necessitates a reconfiguration of classical validity theory so that its core principles can engage with the epistemic conditions introduced by machine learning, where representation, inference, and application are tightly coupled.

At the foundation of this framework lies the requirement that a predicted property be meaningfully aligned with the theoretical construct it claims to represent. The question is not merely whether a model produces numerically accurate outputs, but whether those outputs can be interpreted as instantiations of physically grounded quantities. In the case of bandgap prediction, this entails demonstrating that the model encodes relationships consistent with quantum-mechanical electronic structure rather than relying on statistically convenient compositional proxies. As emphasized in foundational validity theory [2], such alignment demands coherence between the internal logic of the model and established physical understanding. In practice, this involves interrogating feature attributions or attention patterns to determine whether they correspond to known descriptors such as orbital interactions or lattice symmetry, rather than reflecting artifacts of dataset composition. Situations in which a formation-energy predictor appears accurate while relying on elemental prevalence patterns that fail in underrepresented chemistries illustrate how this alignment can break down, thereby undermining the interpretive legitimacy of the output [3].

Once the interpretive grounding is established, the question of external anchoring becomes unavoidable, particularly in a domain where independent measurements are sparse and often costly. The relationship between AI-generated values and external criteria—whether experimental observations or high-fidelity simulations—provides a critical, though incomplete, source of validation. Prior work has shown that correlation with such benchmarks is frequently treated as sufficient evidence of model quality [4], yet this perspective risks conflating statistical agreement with epistemic adequacy. A more rigorous approach requires specifying, in advance, which external criteria are relevant for the intended application and evaluating whether the model supports the same inferential conclusions that those criteria would warrant. This emphasis on inferential equivalence prevents uncritical reliance on proxy datasets that may themselves encode systematic biases, thereby situating criterion-based evaluation within a broader argumentative structure rather than treating it as a standalone metric.

The stability of these arguments under variation in chemical, structural, and operational conditions introduces a further layer of complexity. Materials models are often developed and validated within relatively narrow domains, yet their outputs are routinely extrapolated to unexplored regions of compositional and processing space. Evidence from the literature indicates that performance can degrade sharply under such shifts [5], raising concerns not only about predictive accuracy but also about the persistence of the intended interpretation. Ensuring that a model retains its conceptual meaning across these regimes requires more than empirical testing; it calls for deliberate stress-testing of boundary conditions and reasoned justification for why the learned representation should remain valid when confronted with novel chemistries or extreme environments. Without such scrutiny, predictions that appear reliable within the training manifold may authorize invalid inferences when deployed in inverse-design settings targeting unfamiliar materials systems.

Closely related to this issue is the question of how validity claims respond to perturbations that reflect the realities of materials data and modeling workflows. Variations in input structures, uncertainties in simulation parameters, and shifts in data distributions can all expose latent fragilities in machine-learning models. Prior analyses have highlighted how such vulnerabilities constitute a central challenge in the field [7], yet their implications extend beyond performance degradation. A model may retain low aggregate error under perturbation while losing the capacity to represent the physical mechanisms that define the target property. Robustness, in this sense, must be understood not only as resistance to error inflation but as the preservation of interpretive coherence under conditions of uncertainty and variation.

Even when a model satisfies these interpretive and evidentiary demands, the consequences of its deployment introduce an additional dimension that cannot be ignored. Treating AI-generated outputs as legitimate measurements carries implications for experimental design, resource allocation, and, in some cases, safety-critical decision-making. The broader validity framework developed in measurement theory places particular emphasis on such consequences [1], and this concern becomes especially salient in materials AI, where predictive outputs increasingly guide real-world actions. A model that underestimates fracture toughness, for instance, may appear satisfactory under conventional evaluation yet lead to the selection of unsuitable materials for structural applications, with ramifications that extend beyond the computational domain [8]. Addressing this dimension requires explicit consideration of potential misuses, asymmetries in access to computational resources, and the longer-term effects on experimental validation practices.

Rather than functioning as discrete checkpoints, these dimensions operate through mutual reinforcement and constraint. Weakness in the alignment between model representation and physical construct can propagate through external validation, limit generalizability, and amplify downstream risks. At the same time, strong performance across varying conditions can, in turn, support the credibility of the interpretive argument. The resulting structure is best understood as a tightly coupled network in which the AI-generated property occupies a central position, continuously shaped and evaluated through its relationships to these interdependent dimensions. Within this network, validity emerges not as a checklist of satisfied criteria but as an evolving, evidence-based judgment that integrates theoretical reasoning with empirical performance.

Table 2 integrates the five theoretical components with the five validity dimensions, providing a structured mapping between conceptual arguments, evaluative criteria, and material-specific validity threats.

Table 2. Integrated structure of the five validity components and five validity dimensions

Theoretical component	Associated validity dimension	Key question	Typical threat	Required evidence type	Example in materials AI
Interpretive argument	Construct validity	Does the prediction represent the intended physical construct?	Proxy learning and spurious correlations	Model interpretability and physics alignment	Bandgap predicted via compositional proxy instead of electronic structure
Use argument	Criterion validity	Is the prediction appropriate for the intended decision or application?	Misalignment between training data and the use context	External benchmarks and experimental validation	Formation energy used for synthesis decisions outside equilibrium assumptions
Evidence integration	Generalizability validity	Does validity hold across materials, spaces, and conditions?	Dataset bias and narrow chemical coverage	Cross-domain testing and conceptual stress cases	The model fails when moving from oxides to nitrides
Validity threats	Robustness validity	Does the interpretive claim survive perturbations?	Sensitivity to noise or distribution shift	Perturbation analysis, adversarial testing	Thermal conductivity prediction is unstable under structural noise
Ongoing validation Process	Consequential validity	What are the downstream impacts of using the prediction?	Misuse, overconfidence, and safety risks	Scenario analysis and impact evaluation	Underestimated fracture toughness leading to unsafe material selection

By elaborating each dimension with materials-specific definitions, examples, and evaluation strategies, the theory equips researchers to move beyond unidimensional accuracy reporting toward a multi-faceted validity argument tailored to the complexities of AI-driven property prediction [3].

Relation to Existing Concepts

The proposed theory of measurement validity situates itself in deliberate relation to four concepts that already pervade materials artificial intelligence—accuracy, reliability, reproducibility, and uncertainty quantification—while maintaining that none of these can substitute for validity. Accuracy, understood as closeness of predicted values to reference data, is repeatedly treated as the primary validation criterion in the literature [4, 5, 8]. The present framework explicitly positions accuracy as one component of evidence integration (Component 3) rather than a standalone guarantee of validity; an accurate prediction may still be construct-invalid if it captures a proxy rather than the target physical quantity [1].

Reliability, by contrast, refers to the consistency or repeatability of the measurement under unchanged conditions. In psychometrics, this is distinct from validity, and the same distinction applies here: an AI model may produce highly reproducible outputs across identical inputs yet remain invalid if those outputs misrepresent the intended construct. For example, a neural network that consistently predicts the same erroneous bandgap because of a systematic training bias

exhibits high reliability but zero construct validity [2]. Thus, reliability is necessary—unreliable predictions cannot support any valid inference—but it is far from sufficient.

Reproducibility concerns the ability of independent researchers to obtain the same predictions given the same model and data. While reproducibility is essential for scientific credibility and is addressed in several materials-informatics studies [10, 12], it addresses only the consistency of the computational pipeline, not the epistemic legitimacy of the inferences drawn from the outputs. A fully reproducible but construct-misaligned model still fails the validity test. The theory, therefore, treats reproducibility as a prerequisite that enables validity assessment rather than a replacement for it.

Uncertainty quantification has gained prominence as a means of expressing epistemic doubt around predictions [17-29]. Yet quantifying uncertainty does not automatically confer validity; a model may provide well-calibrated uncertainty estimates while still generating outputs that are inappropriate for the intended use or that carry adverse consequences. The framework, therefore, incorporates uncertainty quantification within evidence integration and validity-threat analysis but insists that it must be interpreted through the lens of the interpretive and use arguments.

By distinguishing validity from these neighboring concepts, the theory clarifies that materials artificial intelligence must move beyond isolated metrics—whether error bars, standard deviations, or reproducibility scores—toward a holistic argument that links model outputs to physical meaning and practical utility [3]. This relational positioning prevents the common conflation that has left much of the field's validation literature conceptually incomplete.

Implications for Materials AI Practice

The conceptual theory carries direct implications for how authors, reviewers, and the broader community should approach validation of AI-generated materials properties. For authors, the framework requires explicit articulation of both the interpretive argument and the use argument before any accuracy metric is reported. Rather than concluding a manuscript with a table of mean absolute errors, authors must now delineate what physical construct their model output is claimed to represent and specify the downstream decisions—such as high-throughput screening or experimental prioritization—that the predictions are intended to support [1]. In addition, authors should present evidence addressing all five validity dimensions and openly discuss validity threats, thereby transforming the validation section from a performance summary into a reasoned scientific defense.

Reviewers, in turn, must shift their evaluation criteria. Instead of asking only whether error metrics fall below community thresholds, reviewers should probe whether the manuscript articulates a coherent validity argument and supplies converging evidence beyond accuracy [3]. Specific questions might include: Does the paper demonstrate that the model captures the intended construct rather than a dataset artifact? Have potential consequential impacts been considered? Is generalizability across relevant chemical spaces addressed? Such reviewer expectations would raise the epistemological standard of the field and discourage publication of models whose validity rests solely on test-set performance [7, 8].

For the materials-informatics community at large, the theory implies the development of standardized validity-reporting templates analogous to those already used in psychometrics. These templates could require authors to address each of the five dimensions and the five theory components in a structured appendix or dedicated section. The community should also invest in creating validity benchmarks that go beyond current accuracy-focused leaderboards; new benchmarks might include curated out-of-distribution test sets designed to probe construct validity or case studies that evaluate consequential validity in simulated decision-making scenarios. Over time, these practices would foster a culture in which validity is treated as an ongoing, collective responsibility rather than an individual model attribute [4, 5].

Collectively, these changes reposition validation as a core scientific activity rather than a procedural formality, ensuring that AI-generated properties earn the right to be treated as trustworthy measurements within materials discovery pipelines.

Conclusion

This conceptual framework has introduced a theory of measurement validity for AI-generated materials properties that adapts foundational ideas from psychometrics and the philosophy of science to the unique challenges of machine-learning inference in the materials domain. By distinguishing accuracy from validity and by articulating five interlocking components—interpretive argument, use argument, evidence integration, validity threats, and validation as an ongoing process—the theory provides a structured yet flexible scaffold for evaluating model outputs. The five dimensions of validity—construct, criterion, generalizability, robustness, and consequential—further operationalize the framework, offering researchers concrete criteria against which to judge whether a predicted property legitimately supports intended scientific or engineering inferences.

The central thesis is clear: a low prediction error does not, by itself, license the use of an AI-generated value in downstream decision-making. Only when accuracy is embedded within a broader validity argument, supported by multiple lines of evidence and tempered by explicit acknowledgment of threats and consequences, can materials artificial intelligence claim to produce measurements worthy of trust. The framework, therefore, calls on the community to move beyond accuracy-centric validation toward validity-aware practice. Authors, reviewers, and collective standards must now prioritize purpose-driven reasoning, ensuring that every AI-generated property is accompanied by a transparent argument linking model output to physical meaning and practical utility.

In doing so, the field will elevate the epistemological rigor of AI-driven materials discovery, transforming predictive models from correlative tools into instruments of genuine scientific insight. The ultimate promise is a materials informatics enterprise in which validity is not an afterthought but the foundational criterion that separates trustworthy knowledge from mere numerical agreement.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 19 Apr 2023 Revised: 31 May 2023 Accepted: 09 Jul 2023
Published online: 18 January 2024

References

Martin-Key NA, Spadaro B, Funnell E, Barker EJ, Schei TS, Tomasik J, et al. The current state and validity of digital assessment tools for psychiatry: Systematic review. JMIR Ment Health. 2022;9(3):e32824.

Trafimow D. Construct validity. Int J Aviat Res. 2022;14(1).

Pillet JC. A linguistic perspective on ai-generated scales: Readability, diversity, and content validity considerations. ECIS; 2024. p. 4. Available from: https://aisel.aisnet.org/ecis2024/track22_innrm/track22_innrm/4

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.

Zunger A. Inverse design in search of materials with target functionalities. Nat Rev Chem. 2018;2(4):0121.

Gubernatis JE, Lookman T. Machine learning in materials science: The possibilities and the pitfalls. MRS Bull. 2018;43(9):648-52. https://doi.org/10.1557/mrs.2018.191.

Li K, DeCost B, Choudhary K, Greenwood M, Hattrick-Simpers J. A critical examination of robustness and generalizability of machine learning prediction of materials properties. NPJ Comput Mater. 2023;9(1):55.

Huang G, Guo Y, Chen Y, Nie Z. Application of machine learning in material synthesis and property prediction. Materials. 2023;16(17):5977.

Schleder GR, Padilha AC, Acosta CM, Costa M, Fazzio A. From DFT to machine learning: Recent approaches to materials science–a review. J Phys Mater. 2019;2(3):032001.

Gossett E, Toher C, Oses C, Isayev O, Legrain F, Rose F, et al. AFLOW-ML: A restful api for machine-learning predictions of materials properties. Comput Mater Sci. 2018;152:134-45.

Kauwe SK, Graser J, Murdock R, Sparks TD. Can machine learning find extraordinary materials? Comput Mater Sci. 2020;174:109498.

Varoquaux G, Colliot O. Evaluating machine learning models and their diagnostic value. In: Machine learning for brain disorders. 2023. p. 601-30.

Taufique MF, Mamun O, Roy A, Khakurel H, Balasubramanian G, Ouyang G, et al. Machine learning guided prediction of the yield strength and hardness of multi-principal element alloys. Mater Open Res. 2024;2:9.

Qayyum F, Khan MA, Kim DH, Ko H, Ryu GA. Explainable AI for material property prediction based on energy cloud: A Shapley-driven approach. Materials. 2023;16(23):7322.

Hu J, Liu D, Fu N, Dong R. Realistic material property prediction using domain adaptation based machine learning. Digit Discov. 2024;3(2):300-12.

Li Q, Fu N, Omee SS, Hu J. MD-HIT: Machine learning for material property prediction with dataset redundancy control. NPJ Comput Mater. 2024;10(1):245.

Li Z, Li S, Birbilis NJ. A machine learning-driven framework for the property prediction and generative design of multiple principal element alloys. Mater Today Commun. 2024;38:107940.

Reich Y, Barai SV. Evaluating machine learning models for engineering problems. Artif Intell Eng. 1999;13(3):257-72.

Tavazza F, DeCost B, Choudhary K. Uncertainty prediction for machine learning models of material properties. ACS Omega. 2021;6(48):32431-40.

Kiyasseh D, Cohen A, Jiang C, Altieri N. SUDO: A framework for evaluating clinical artificial intelligence systems without ground-truth annotations. arXiv [Preprint]. 2024 Jan 2:arXiv:2403.17011.

Laurer M, Van Atteveldt W, Casas A, Welbers K. On measurement validity and language models: Increasing validity and robustness against bias with instructions [Internet]. 2023.

Jha D, Gupta V, Ward L, Yang Z, Wolverton C, Foster I, et al. Enabling deeper learning on big data for materials informatics applications. Sci Rep. 2021;11(1):4244.

Yang J, Li FZ, Arnold FH. Opportunities and challenges for machine learning-assisted enzyme engineering. ACS Cent Sci. 2024;10(2):226-41.

Mileo PG, Cho KH, Chang JS, Maurin G. Water adsorption fingerprinting of structural defects/capping functions in Zr–fumarate MOFs: A hybrid computational-experimental approach. Dalton Trans. 2021;50(4):1324-33.

Lebovitz S, Levina N, Lifshitz-Assaf H. Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. MIS Q. 2021;45(3):1501-26.

Kim M, Saad W, Mozaffari M, Debbah M. On the tradeoff between energy, precision, and accuracy in federated quantized neural networks. In: ICC 2022-IEEE International Conference on Communications. IEEE; 2022. p. 2194-9.

Li J, Liu L, Le TD, Liu J. Accurate data-driven prediction does not mean high reproducibility. Nat Mach Intell. 2020;2(1):13-5.

Van Noorden R, Perkel JM. AI and science: What 1,600 researchers think. Nature. 2023;621(7980):672-5.