ORIGINAL RESEARCH **Open access**

Scientific Blind Spots Introduced by Feature Engineering in Materials Informatics

Fatima Zahra Amrani[1*], Youssef Benali[1], Samira El-Haddad[2]

Abstract

Feature engineering remains central to materials informatics, yet systematically introduces scientific blind spots that constrain discovery and interpretation. These blind spots arise from choices in descriptor selection, transformation, and dimensionality reduction that inadvertently prioritize statistical correlations over physical invariance, overlook multi-scale interactions, and embed dataset-specific biases into model architectures. In small-data regimes common to materials science, engineered features often amplify overfitting while diminishing generalizability across chemical spaces. Interpretability suffers as complex engineered descriptors obscure mechanistic linkages between atomic structure and macroscopic properties. Literature consistently highlights these limitations across perovskites, alloys, energy materials, and porous systems, underscoring the tension between predictive performance and scientific fidelity. This conceptual manuscript synthesizes these challenges and proposes an original Integrated Blind Spot Navigation Model (IBSNM). The framework organizes feature engineering around four interdependent pillars—physical consistency guardrails, multi-scale descriptor integration, uncertainty-aware selection, and iterative co-interpretation—linked by feedback mechanisms that surface and mitigate hidden assumptions. By reframing feature engineering as a navigable landscape rather than a static preprocessing step, the model offers a conceptual pathway toward more robust, transparent materials informatics practices that do not rely on empirical validation.

Keywords Materials informatics, Interpretability, Feature engineering, Scientific blind spots, Descriptor selection, Generalizability

*Correspondence:
Fatima Zahra Amrani
fatima.amrani@gmail.com
[1] Department of Computational Materials Science, Faculty of Engineering, Mohammed V University, Rabat, Morocco
[2] Department of Artificial Intelligence in Engineering Systems, Faculty of Engineering, University of Fez, Fez, Morocco

Introduction

Materials informatics has transformed the discovery of new compounds by leveraging machine learning on large descriptor spaces derived from atomic, electronic, and structural data [1-6]. Feature engineering—the deliberate crafting and selection of input variables—underpins these advances but simultaneously generates scientific blind spots—systematic omissions or distortions in understanding that arise from representational choices rather than inherent data scarcity. These blind spots manifest as loss of physical interpretability, reduced transferability across compositional domains, and unintended amplification of spurious correlations [3, 7-11].

Early enthusiasm for high-dimensional descriptors gave way to recognition that engineered features can embed assumptions about symmetry, locality, and scale that conflict with the complexity of real materials [7, 12]. For instance, fixed topological or geometric descriptors may fail to capture emergent phenomena in disordered systems or under varying thermodynamic conditions. In small-data environments typical of experimental materials datasets, aggressive

feature reduction further exacerbates these issues by discarding higher-order interactions essential for property prediction [13, 14].

The literature reveals recurring themes: descriptor instability across datasets, trade-offs between model accuracy and mechanistic insight, and challenges in reconciling data-driven representations with established physical principles [4, 15-17]. Studies on perovskites and battery materials illustrate how feature choices can mask compositional sensitivities or introduce artificial biases toward well-represented classes [16, 18]. Despite progress in explainable methods, the fundamental question persists: how do engineered features shape—not merely represent—the scientific questions posed in materials informatics?

This conceptual analysis examines these blind spots through interpretive synthesis rather than empirical evaluation. It draws on peer-reviewed contributions published strictly to map the conceptual terrain. The manuscript first synthesizes theoretical foundations and literature insights, then articulates an original conceptual framework designed to navigate rather than eliminate these blind spots. The proposed model reframes feature engineering as a dynamic, reflexive process that acknowledges inherent limitations while fostering iterative refinement of representational assumptions. In doing so, it seeks to advance a more reflexive practice in materials informatics, one attuned to the epistemic constraints imposed by human-designed descriptors.

Theoretical Background and Literature Synthesis

The role of feature engineering in materials informatics: Feature engineering transforms raw structural and compositional data into informative descriptors suitable for machine learning models [1, 2, 4]. Foundational works emphasize workflows that balance expressiveness with computational tractability, often incorporating domain knowledge through hand-crafted invariants such as coordination numbers, bond lengths, and electronic fingerprints [5-7]. These engineered representations enable prediction of formation energies, band gaps, and mechanical properties across diverse material classes. However, the process inherently involves selection biases that prioritize certain scales or symmetries, potentially sidelining alternative physical perspectives [8, 10].

Emergence of blind spots in descriptor selection: A recurring conceptual concern is the introduction of representational blind spots through dimensionality reduction and feature filtering [3, 11, 12]. When descriptors are chosen to maximize statistical performance on limited datasets, they may capture dataset-specific artifacts rather than invariant physical relationships [13, 14]. The literature highlights that compressed-sensing approaches and automated selection tools can overlook subtle multi-body interactions that are critical for phase stability or defect behavior [15, 16]. In alloy and perovskite systems, engineered features frequently fail to generalize when compositional complexity increases, revealing hidden assumptions about local ordering [17, 18]. A consolidated typological mapping of feature-engineering-induced blind spots, including their epistemic origins and materials-specific manifestations, is presented in **Table 1**.

Table 1. Typological mapping of scientific blind spots emerging from feature engineering practices in materials informatics

Blind spot category	Feature engineering origin	Mechanism of epistemic distortion	Manifestation in material contexts	Conceptual consequence
Physical invariance violations	Descriptor normalization, coordinate encoding	Breaks symmetry, conservation, or invariance principles	Misrepresentation of crystal symmetry, rotational artifacts	Physically incoherent predictions
Scale isolation bias	Single-scale descriptor construction	Neglects cross-scale coupling effects	Failure to capture defect propagation or grain interactions	Loss of emergent phenomena visibility

Descriptor compression loss	Dimensionality reduction, feature filtering	Removes higher-order interactions	Oversimplified phase stability modeling	Mechanistic underrepresentation
Dataset imprinting bias	Training-set-driven descriptor optimization	Encodes statistical artifacts as physical signals	Overfitting to dominant chemistries (e.g., oxides)	Reduced transferability
Multi-body interaction blindness	Fixed geometric/topological descriptors	Underrepresents nonlocal interactions	Poor modeling of alloy disorder or defect energetics	Incomplete structural causality
Interpretability obfuscation	Composite engineered descriptors	Masks causal atomic–property linkages	Opaque feature importance maps	Reduced mechanistic insight
Small-data amplification effects	Feature selection under sparse data	Inflates noise-driven correlations	Instability across validation datasets	Fragile generalizability
Thermodynamic context omission	Static structural descriptors	Ignores environmental conditions	Misleading stability predictions	Contextual incompleteness

Interpretability challenges arising from engineered features: Explainable machine learning techniques have illuminated how opaque engineered descriptors hinder causal inference [3, 7, 11]. Complex transformations can obscure direct links between atomic configurations and emergent properties, leading to models whose predictions lack mechanistic grounding [19, 20]. Reviews stress that while graph-based or learned representations reduce reliance on manual engineering, hybrid approaches still confront trade-offs between predictive power and human-interpretable insights [21, 22].

Generalizability and small-data constraints: In regimes with sparse experimental data, feature engineering amplifies risks of overfitting and poor extrapolation [6, 8, 10, 13]. Studies on energy materials and solid-state compounds demonstrate that descriptors optimized for one chemical family often introduce systematic errors when applied to chemically dissimilar systems [14-16]. Uncertainty quantification remains underdeveloped relative to feature generation, leaving blind spots unaddressed [23, 24].

Synthesis of current understanding: Collective literature portrays feature engineering as a double-edged process: indispensable for enabling data-driven discovery yet generative of epistemic blind spots through implicit assumptions about invariance, scale, and relevance [1]. These works converge on the need for reflexive frameworks that treat feature choices as hypotheses open to continuous interrogation rather than fixed inputs. The synthesis reveals that addressing blind spots conceptually requires integrating physical priors, multi-scale perspectives, and iterative feedback without presuming empirical resolution.

Proposed conceptual framework

The Integrated Blind Spot Navigation Model (IBSNM)

The Integrated Blind Spot Navigation Model (IBSNM) advances a systems-level conceptual architecture for navigating scientific blind spots embedded within feature engineering practices in materials informatics. Rather than framing feature engineering as a deterministic preprocessing stage within linear machine learning pipelines, IBSNM repositions it as an epistemic navigation landscape—dynamic, reflexive, and structurally conditioned by both physical knowledge and algorithmic abstraction. Within this interpretive lens, blind spots are not treated as anomalous oversights but as structurally emergent phenomena arising from representational compression, scale discontinuities, descriptor biases, and implicit modeling assumptions.

IBSNM is organized around four interdependent conceptual pillars linked through recursive feedback circuits. These pillars function not as modular pipeline stages but as coevolving interpretive strata that collectively shape the legitimacy, robustness, and scientific expressivity of descriptors.

Pillar I — Physical consistency guardrails

The first pillar, physical consistency guardrails, establishes boundary conditions that constrain the construction of descriptors within physically meaningful representational spaces. In materials informatics, feature engineering often abstracts atomic and electronic structures into numerical descriptors optimized for predictive performance. However, such abstractions risk violating invariance principles, conservation laws, or symmetry operations fundamental to materials physics.

Within IBSNM, guardrails operate as conceptual sentinels that interrogate descriptors against known physical constraints. These include rotational and translational invariance, preservation of point-group symmetry, thermodynamic feasibility bounds, and stoichiometric consistency. When descriptors transgress these constraints—whether through improper normalization, scale distortion, or latent-feature entanglement—the guardrail system triggers a reflexive re-evaluation.

Importantly, this pillar is not prescriptive; it does not eliminate unconventional descriptors outright. Instead, it categorizes violations along interpretive gradients ranging from benign abstraction to epistemically destabilizing distortion. In doing so, Physical Consistency Guardrails preserve exploratory flexibility while preventing representational drift into physically incoherent regimes.

Illustrative sub-elements include:

- Symmetry invariants and equivariant embeddings
- Thermodynamic and kinetic feasibility bounds
- Conservation-law encoding
- Physically constrained normalization schemas

Pillar II — Multi-scale descriptor integration

The second pillar, multi-scale descriptor integration, addresses representational discontinuities that arise when material phenomena are encoded at isolated spatial or temporal scales. Materials behavior emerges through cross-scale interactions spanning electronic orbitals, atomic bonding networks, microstructural topology, and continuum mechanics. Feature engineering strategies that privilege a single scale risk, obscuring emergent phenomena such as defect propagation, phase coupling, or hierarchical reinforcement mechanisms.

IBSNM conceptualizes descriptor construction as a vertically layered representational stack integrating atomic, mesoscopic, and continuum features. Crucially, this pillar emphasizes cross-scale reconciliation mechanisms rather than mere feature concatenation. Reconciliation processes interrogate scale compatibility and resolve descriptor conflicts through weighting logics, hierarchical embeddings, or scale-bridging latent representations.

This pillar also recognizes epistemic asymmetry across scales: atomic descriptors often exhibit high precision but limited contextual scope, whereas continuum descriptors provide systems context but may dilute atomistic causality. Multi-Scale Descriptor Integration, therefore, functions as a balancing architecture that preserves fine-grained physical fidelity while enabling systems-level interpretability.

Illustrative sub-elements include:

- Atomic graph embeddings and orbital descriptors
- Grain boundary and defect topology metrics

- Microstructural morphology encodings
- Continuum thermomechanical property fields

Pillar III — Uncertainty-aware feature selection

The third pillar, uncertainty-aware feature selection, embeds probabilistic reasoning directly within descriptor evaluation. Conventional feature selection techniques prioritize statistical relevance or predictive contribution but often neglect epistemic fragility—how descriptor reliability fluctuates under chemical perturbation, compositional extrapolation, or data scarcity.

Within IBSNM, descriptors are evaluated not only for predictive salience but for stability across hypothetical perturbation landscapes. This includes simulated compositional substitutions, structural distortions, and temperature-pressure variations. Descriptors demonstrating volatility under such perturbations are flagged as epistemically fragile, prompting recalibration or redundancy pairing.

Uncertainty is conceptualized across multiple strata:

- Aleatoric uncertainty from measurement noise
- Epistemic uncertainty from sparse training regimes
- Representational uncertainty from descriptor abstraction

By embedding uncertainty diagnostics into feature selection, this pillar transforms static predictors into probabilistically annotated knowledge carriers.

Illustrative sub-elements include:

- Descriptor sensitivity mapping
- Bayesian feature robustness scoring
- Perturbation response simulations
- Confidence-weighted feature ranking

Pillar IV — Iterative human–machine co-interpretation

The fourth pillar, iterative human–machine co-interpretation, repositions expert judgment as an ongoing interpretive partner rather than a front-loaded supervisory input. Feature engineering decisions often encode tacit disciplinary assumptions—choices about which physical phenomena matter, which scales dominate, and which simplifications are acceptable.

IBSNM frames human expertise and machine inference as dialogic agents engaged in recursive interpretive exchange. Algorithms surface latent feature correlations, anomalies, or representational compressions, while domain experts interrogate these outputs through theoretical, experimental, and phenomenological lenses.

This co-interpretive loop surfaces blind spots that neither agent could independently detect. For example, an algorithm may identify a high-salience descriptor cluster that domain experts recognize as an artifact of dataset bias or measurement coupling.

Illustrative sub-elements include:

- Expert annotation overlays on latent spaces
- Interpretability-guided descriptor pruning
- Hypothesis elicitation from feature clusters

- Reflexive audit trails of feature decisions

Reflexive feedback architecture

A defining feature of IBSNM is its reflexive feedback topology. Each pillar is bidirectionally coupled with the others, producing iterative recalibration cycles:

- Guardrail violations inform uncertainty scoring
- Cross-scale conflicts trigger co-interpretive review
- Expert critiques reshape physical constraint encoding
- Uncertainty diagnostics prompt descriptor reintegration

These recursive exchanges transform blind spot navigation into a continuous epistemic monitoring process rather than a one-time validation step.

Circumferential cycling mechanisms further reinforce this dynamism, institutionalizing periodic re-evaluation as datasets evolve, models retrain, and scientific knowledge expands. The navigational functions through which IBSNM operationalizes blind spot exposure, mitigation, and documentation across its four pillars and reflexive couplings are synthesized in **Table 2.**

Table 2. Navigational functions of the integrated blind spot navigation model (IBSNM) in mitigating feature engineering blind spots

IBSNM pillar	Blind spot targets	Navigational mechanism	Feedback inputs	Epistemic outcome
Physical consistency guardrails	Invariance violations, thermodynamic omission	Constraint-based descriptor auditing	Expert review, symmetry diagnostics	Physically coherent feature spaces
Multi-scale descriptor integration	Scale isolation, emergent interaction blindness	Cross-scale descriptor layering	Structural–continuum reconciliation signals	Restored hierarchical interpretability
Uncertainty-aware feature selection	Dataset imprinting, small-data overfitting	Probabilistic robustness scoring	Perturbation simulations, variance metrics	Descriptor reliability transparency
Human–machine co-interpretation	Interpretability loss, latent bias embedding	Reflexive dialogic feature evaluation	Expert annotation, explainability outputs	Surfaced tacit assumptions
Guardrail ↔ Uncertainty coupling	Descriptor instability	Violation-triggered uncertainty scoring	Constraint deviation signals	Risk-calibrated selection
Scale ↔ Co-interpretation Coupling	Cross-scale conflict opacity	Expert arbitration of descriptor tensions	Interpretive audits	Resolved scale discordance
Uncertainty ↔ Integration coupling	Perturbation-sensitive scale features	Stability-informed reintegration	Sensitivity gradients	Robust multi-scale encoding
System-level reflexive cycling	Residual blind spots	Iterative recalibration loops	All pillar outputs	Continuous blind spot navigation

Epistemic positioning: Navigating, not eliminating blind spots

IBSNM is grounded in the recognition that complete elimination of the blind spot is conceptually unattainable. Materials systems are characterized by ontological complexity, incomplete observability, and evolving theoretical frameworks. Consequently, the model prioritizes systematic navigation, exposure, and documentation of blind spots over their presumed eradication.

Residual uncertainties are treated as knowledge artifacts—annotated, tracked, and contextualized—thereby preserving transparency in downstream predictive and design decisions. **Figure 1** presents the IBSNM as a circular, integrative framework, in which four synergistic methodological pillars dynamically interact via direct feedback and cyclical refinement to map and navigate a central target property landscape.

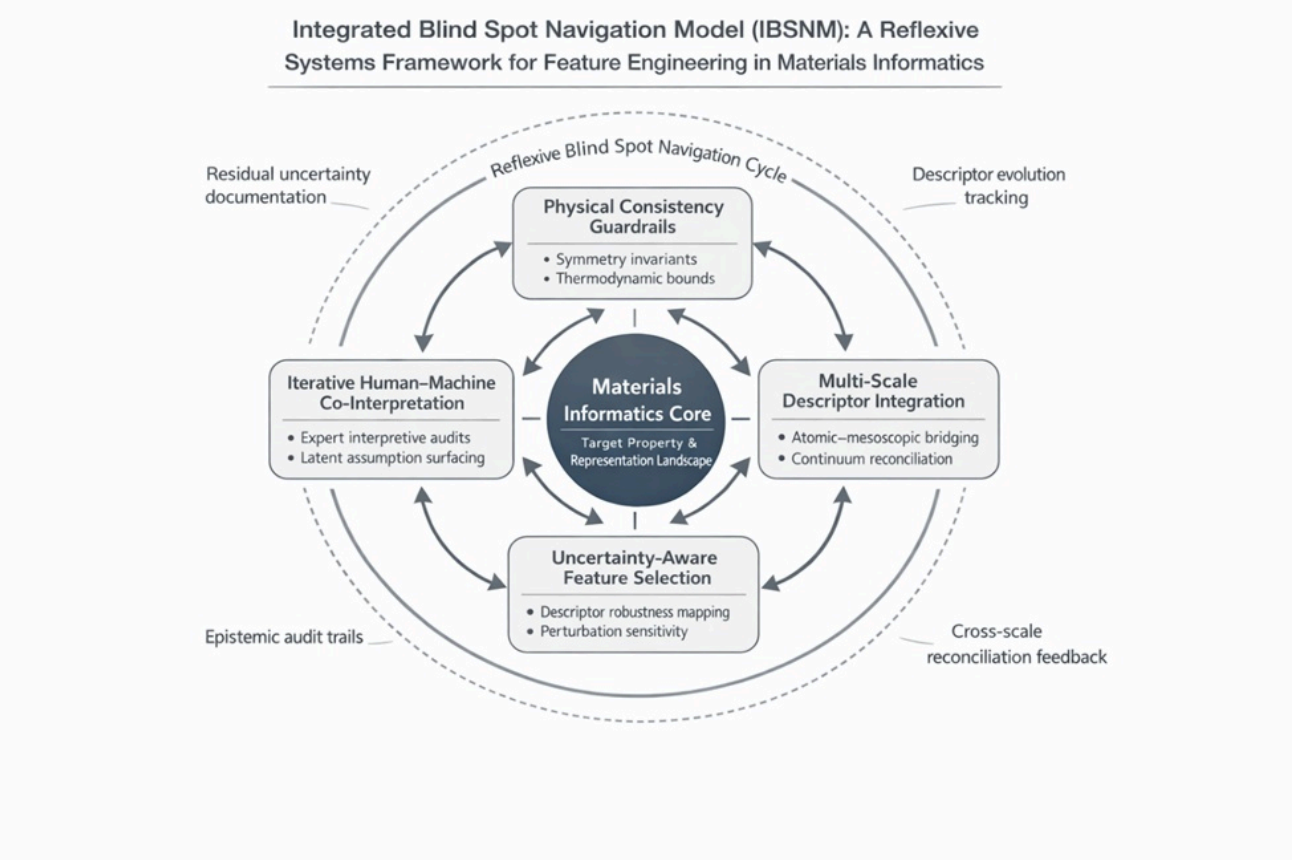

Figure 1. Circular schematic of the integrated bridging scheme for novelty mapping (IBSNM), illustrating the dynamic integration of four synergistic methodological pillars—physical guardrails, multi-scale descriptors, uncertainty-aware selection, and human-machine co-interpretation—around a central materials informatics core.

Propositions

The analytical synthesis of peer-reviewed literature reveals several interpretive propositions concerning the epistemic constraints imposed by feature engineering in materials informatics.

Proposition 1: Descriptor selection processes frequently embed dataset-specific statistical patterns that masquerade as physical invariants, resulting in diminished generalizability when models encounter compositional variations beyond the training domain [1, 3].

Proposition 2: The reliance on fixed geometric and topological descriptors introduces representational blind spots by underrepresenting higher-order or emergent multi-body interactions that govern phase stability and defect dynamics in complex systems [4, 11].

Proposition 3: In small-data regimes, dimensionality reduction techniques amplify overfitting tendencies while simultaneously obscuring mechanistic linkages between atomic configurations and macroscopic properties [1].

Proposition 4: Interpretability tools applied post hoc to engineered features often fail to recover causal physical insights because the initial transformations already encode opaque assumptions about symmetry and locality [3, 11].

Proposition 5: The tension between predictive accuracy and scientific fidelity arises fundamentally from the human-imposed structure of descriptors, which prioritize computational tractability over exhaustive multi-scale consistency [8, 16].

Proposition 6: Reflexive integration of uncertainty quantification with feature selection conceptually surfaces latent biases but cannot fully resolve the inherent incompleteness of any finite descriptor set [17]. These propositions distill recurring conceptual patterns without advancing testable claims, emphasizing the interpretive necessity of acknowledging representational trade-offs.

Results and Discussion

The interpretive propositions above illuminate the pervasive yet under-theorized role of feature engineering in generating scientific blind spots within materials informatics. When descriptors are engineered to maximize statistical fidelity on available datasets, they inherently constrain the scope of inquiry by foregrounding certain invariances—such as local coordination environments—while backgrounding others, including long-range electronic correlations or thermodynamic boundary conditions [1-17]. This selective emphasis creates systematic distortions: models may excel at interpolating within familiar chemical spaces yet falter conceptually when extrapolated, as the engineered representations fail to preserve essential physical symmetries across domains [5, 6].

The literature consistently underscores that small-data constraints exacerbate these issues, as aggressive feature filtering discards subtle signals critical for capturing emergent behaviors in perovskites, alloys, and energy storage compounds [1, 8, 13-16]. The resulting models risk conflating correlation with invariance, leading to predictions that lack robust mechanistic grounding even when augmented by explainability methods [3, 11, 12, 17]. Multi-scale integration offers a partial conceptual remedy by layering descriptors. Yet, it simultaneously introduces new reconciliation challenges, as atomic-level features may conflict with continuum approximations in ways that remain opaque without reflexive interrogation [4, 11, 14].

Uncertainty-aware selection and human-machine co-interpretation, as conceptualized in the IBSNM framework, provide navigational structures rather than solutions. They encourage ongoing documentation of residual blind spots—such as unmodeled compositional sensitivities or scale mismatches—without presuming elimination of epistemic limits [1-17]. Overall, feature engineering functions less as neutral preprocessing and more as a formative epistemic filter that shapes the questions materials informatics can meaningfully address. This perspective calls for sustained analytical vigilance to maintain scientific integrity amid data-driven practices.

Conclusion

Feature engineering occupies a foundational yet ambivalent position in materials informatics: indispensable for enabling computational discovery while systematically generating blind spots through representational choices that prioritize certain invariances over others. The interpretive synthesis presented here, grounded in peer-reviewed scholarship,

underscores the necessity of treating descriptor construction as an inherently reflexive process rather than a technical preliminary step. By centering physical consistency, multi-scale integration, uncertainty awareness, and co-interpretation within the Integrated Blind Spot Navigation Model, the framework offers a conceptual scaffold for navigating—rather than eradicating—these epistemic constraints. Future conceptual development in the field should continue to interrogate the assumptions embedded in engineered features, fostering practices that balance predictive utility with transparent acknowledgment of inherent limitations. This approach ultimately supports a more mature, self-aware materials informatics enterprise attuned to the interpretive complexities of representing complex matter.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 05 Aug 2022 Revised: 18 Sep 2022 Accepted: 13 Nov 2022
Published online: 18 January 2023

References

Wang Y, Zhang Y, Chen C. Small data machine learning in materials science. npj Comput Mater. 2023;9(1):42. https://doi.org/10.1038/s41524-023-01000-z.

Ma Y, Xu P, Li M, Ji X, Zhao W, Lu W. The mastery of details in the workflow of materials machine learning. npj Comput Mater. 2024;10(1):141. https://doi.org/10.1038/s41524-024-01331-5.

Kailkhura B, Gallagher B, Kim S, Hiszpanski A. Explainable machine learning in materials science. npj Comput Mater. 2022;8(1):204.
https://doi.org/10.1038/s41524-022-00884-7.

Wang Z, Chen A, Tao K, Cai J, Gao J, Ye S, et al. AlphaMat: a material informatics hub connecting data, features, models and applications. npj Comput Mater. 2023;9(1):130.
https://doi.org/10.1038/s41524-023-01086-5.

Himanen L, Geurts A, Foster AS, Rinke P. Data-driven materials science: status, challenges, and perspectives. Adv Sci. 2020;7(20):1900808.
https://doi.org/10.1002/advs.201900808.

Morgan D, Jacobs R. Opportunities and challenges for machine learning in materials science. Annu Rev Mater Res. 2020;50:71-103.
https://doi.org/10.1146/annurev-matsci-070218-010015.

Pankajakshan P, Sanyal S, de Albuquerque Oliveira O, Blaiszik B, Foster I, Assary RS, et al. Machine learning for materials scientists: An introductory guide toward best practices. Chem Mater. 2020;32(10):4360-72.
https://doi.org/10.1021/acs.chemmater.0c01907.

Chen C, Zuo Y, Ye W, Li X, Deng Z, Ong SP. A critical review of machine learning of energy materials. Adv Energy Mater. 2020;10(31):1903242.
https://doi.org/10.1002/aenm.201903242.

Zhang Y, Xu L, Berlinguette CP. Feature selection for machine learning-driven accelerated discovery and optimization in emerging photovoltaics: A review. Adv Intell Discov. 2025;2:e202500022.
https://doi.org/10.1002/aidi.202500022.

Jacobs R, Mayeshiba T, Afflerbach B, Miles L, Williams M, Turner M, et al. The materials simulation toolkit for machine learning (MAST-ML): An automated open source toolkit to accelerate data-driven materials research. Comput Mater Sci. 2020;176:109544.
https://doi.org/10.1016/j.commatsci.2020.109544.

Oviedo F, Ferres JL, Buonassisi T, Butler KT. Interpretable and explainable machine learning for materials science and chemistry. Acc Mater Res. 2022;3(6):597-607.
https://doi.org/10.1021/accountsmr.1c00244.

Zhu L, Zhou J, Sun Z. Materials data toward machine learning: advances and challenges. J Phys Chem Lett. 2022;13(18):3965-77.
https://doi.org/10.1021/acs.jpclett.2c00576.

Sendek AD, Cheon G, Pasta M, Reed EJ. Quantifying the search space for high-throughput Li-ion conductor discovery. Chem Mater. 2020;32(1):253-61.
https://doi.org/10.1021/acs.chemmater.9b04078.

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.
https://doi.org/10.1038/s41524-023-01012-9.

Ma B, Wu X, Zhao C, Lin C, Gao M, Sa B, et al. An interpretable machine learning strategy for pursuing high piezoelectric coefficient in Pb-free ceramics. npj Comput Mater. 2023;9:1-12.
https://doi.org/10.1038/s41524-023-01187-1.

Liu X, Xu P, Zhao J, Lu W, Li M, Wang G. Material machine learning for alloys: applications, challenges and perspectives. J Mater Sci Technol. 2023;95:1-18.

Stuart S, Watchorn J, Gu F. Sizing up feature descriptors for macromolecular machine learning with polymeric biomaterials. npj Comput Mater. 2023;9:1-12.
https://doi.org/10.1038/s41524-023-01040-5.

Owens CB, Mathew N, Olaveson TW, Tavenner JP, Kober E, Tucker GJ, et al. Feature engineering descriptors, transforms, and machine learning for grain boundaries and variable-sized atom clusters. npj Comput Mater. 2024;10.
https://doi.org/10.1038/s41524-024-01509-x.

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

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

Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C. Machine learning in materials informatics: recent applications and prospects. npj Comput Mater. 2017;3(1).

Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. Chem Mater. 2019;31(9).

Batra R, Pilania G, Uberuaga BP, Ramprasad R. Multifidelity information fusion with machine learning: a case study of dopant formation energies in hafnia. ACS Appl Mater Interfaces. 2019;11:24906-18.

Wen C, Zhang Y, Wang C, et al. Machine learning assisted design of high entropy alloys with desired property. Acta Mater. 2019;170:109-17.