ORIGINAL RESEARCH **Open access**

Uncertainty as Infrastructure: Governing Confidence in Data-Driven Materials Pipelines

Natalia Petrova[1*], Elena Stoyanova[1], Ivan Dimitrov[2]

Abstract

The field of computational and data-driven materials engineering has transformed traditional discovery processes through the integration of machine learning, high-throughput computations, and autonomous systems. However, as these pipelines scale, the management of uncertainty emerges as a foundational infrastructure rather than a mere analytical byproduct. This manuscript conceptualizes uncertainty not as an obstacle but as an enabling framework for governing confidence in materials informatics workflows. By synthesizing recent advancements in representation learning, graph neural networks, and uncertainty quantification, we identify epistemic gaps in current data-driven ecosystems, where confidence in predictions often remains opaque or inadequately integrated into discovery loops. We introduce the Confidence Governance Framework (CGF), a layered conceptual architecture that embeds uncertainty quantification as a core infrastructural element, facilitating dynamic interactions between data representations, model inferences, and discovery steering. This framework emphasizes computational trade-offs in multimodal datasets and simulation-experiment couplings, promoting robust, interpretable pipelines. Implications extend to enhanced autonomy in inverse design and closed-loop experimentation, fostering resilient materials engineering paradigms. Through this lens, uncertainty becomes a strategic asset for calibrating epistemic risks and optimizing resource allocation in AI-assisted materials research.

Keywords Materials informatics, Uncertainty quantification, Graph neural networks, Representation learning, Computational discovery, Data-driven pipelines

*Correspondence:
Natalia Petrova
natalia.petrova@outlook.com
[1] Department of Computational Materials Engineering, Faculty of Engineering, University of Sofia, Sofia, Bulgaria
[2] Department of Data-Driven Materials Systems, Faculty of Engineering, Technical University of Sofia, Sofia, Bulgaria

Introduction

The evolution of materials engineering has been profoundly shaped by computational advancements, shifting from empirical trial-and-error methods to sophisticated data-driven paradigms. Over the past decade, the integration of machine learning and high-throughput computing has accelerated the exploration of vast materials spaces, enabling the prediction of properties, design of novel compounds, and optimization of synthesis routes [1, 2]. This transition is evident in the proliferation of materials informatics, where large-scale datasets from simulations and experiments fuel predictive models, reducing the time and cost associated with traditional laboratory workflows [3, 4]. Central to this shift is the role of AI ecosystems, which leverage deep learning architectures to process complex structural and compositional data, often represented through graphs or embeddings that capture atomic interactions and symmetries [5, 6].

High-throughput infrastructures have become indispensable, allowing for the systematic screening of thousands of candidates in silico before experimental validation [7, 8]. For instance, computational protocols combining density functional theory with machine learning surrogates have facilitated the discovery of catalysts,

perovskites, and two-dimensional materials by navigating high-dimensional parameter spaces efficiently [9, 10]. Yet, as these systems grow in complexity, they introduce new challenges related to data quality, model generalizability, and the coupling of computational predictions with real-world experimentation [11, 12]. Autonomous discovery systems, which incorporate closed-loop feedback between prediction and synthesis, further amplify these demands, requiring seamless integration across modalities [13, 14].

Despite these progresses, current discovery models exhibit limitations in handling epistemic and aleatoric uncertainties inherent to materials data. Epistemic constraints arise from incomplete knowledge of underlying physics, such as approximations in interatomic potentials or biases in training datasets, while aleatoric factors stem from stochastic processes in simulations and measurements [15, 16]. In high-stakes applications like energy storage or structural alloys, unchecked uncertainties can lead to overconfident predictions, misallocated resources, or failed translations from computation to fabrication [17, 18]. Moreover, the opacity of black-box models, particularly in graph neural networks, complicates interpretability, hindering the traceability of confidence in pipeline outputs [19, 20].

Computational design paradigms, including inverse materials design, rely on robust representations to invert property-to-structure mappings, but often overlook the infrastructural role of uncertainty in governing these inversions [21, 22]. For example, while foundation models for science promise generalizable embeddings across materials classes, their deployment in pipelines necessitates mechanisms to quantify and propagate confidence, ensuring that discovery steering aligns with epistemic realities [23, 24]. The literature highlights a need for infrastructures that treat uncertainty as a core component, rather than an afterthought, to enhance the reliability of AI-guided systems [25, 26].

This manuscript addresses these gaps by framing uncertainty as an infrastructural element in data-driven materials pipelines. We synthesize key developments in the field to underscore the interplay between data infrastructures, learning architectures, and discovery dynamics. Ultimately, we position a novel conceptual framework that integrates uncertainty quantification into the fabric of computational workflows, enabling governed confidence that steers materials engineering toward more resilient and efficient outcomes.

Theoretical Background & Literature Synthesis

Materials data infrastructures

Materials data infrastructures form the backbone of computational discovery, encompassing the curation, integration, and utilization of multimodal datasets from simulations, experiments, and literature. High-throughput computing has enabled the generation of extensive repositories, such as those cataloging electronic structures, thermodynamic properties, and mechanical behaviors across elemental and compound spaces [7, 9]. These infrastructures support the scaling of data-driven approaches, where standardized formats facilitate interoperability between computational tools and experimental validations [13, 17]. However, the heterogeneity of data sources—ranging from ab initio calculations to spectroscopic measurements—poses challenges in maintaining consistency and addressing gaps in coverage [24, 27].

In this context, multimodal datasets emerge as critical, combining structural, compositional, and functional attributes to inform comprehensive models [22, 25]. The infrastructure must account for data provenance, ensuring traceability from raw inputs to derived insights, which is essential for mitigating biases that propagate through pipelines [11, 28]. Literature emphasizes the need for robust data management systems that enable dynamic querying and augmentation, supporting the iterative refinement of materials knowledge bases [8, 14].

Representation learning architectures

Representation learning has revolutionized how materials are encoded for computational analysis, with graph neural networks (GNNs) and deep architectures providing expressive embeddings that capture atomic connectivity and symmetries [3-6]. These models transform raw structural data into latent spaces amenable to property prediction and optimization, outperforming traditional descriptors in tasks like bandgap estimation or catalyst screening [10, 18, 19]. Advances in atomistic line graphs and message-passing schemes enhance the interpretability of these representations, allowing for the disentanglement of local and global features [5, 29].

Yet, the choice of architecture influences the fidelity of representations, particularly in polycrystalline or disordered

systems where multi-scale interactions must be resolved [3, 30]. Foundation models extend this by pre-training on vast corpora, enabling transfer learning across materials domains and reducing the need for domain-specific tuning [20, 21]. The synthesis reveals a trend toward hybrid architectures that incorporate physical priors, such as invariance to rotations and translations, to bolster generalization in data-scarce regimes [1, 2, 12].

AI-Guided discovery systems

AI-guided systems integrate machine learning with autonomous workflows, orchestrating closed-loop experimentation where predictions inform synthesis and vice versa [7, 13, 19]. Active learning strategies, such as those employing neural networks for alloy development, optimize exploration by selecting informative candidates, thereby accelerating convergence on optimal materials [19, 31]. High-throughput screening protocols exemplify this, combining computational surrogates with experimental feedback to discover bimetallic catalysts or transparent conductors [7, 9].

The dynamics of these systems hinge on seamless simulation-experiment coupling, where discrepancies between predicted and measured properties drive model updates [14, 15]. Literature underscores the role of inverse design in these paradigms, where generative models map desired properties back to viable structures, navigating combinatorial spaces efficiently [21, 29]. However, the efficacy of such systems depends on infrastructural support for real-time data assimilation and error propagation [16, 26].

Computational design paradigms

Computational design paradigms encompass inverse approaches, where optimization algorithms steer toward target functionalities, often leveraging multi-fidelity models to balance accuracy and efficiency [11, 27, 30]. High-throughput methods for 2D materials or interfaces illustrate this, generating potential energy surfaces or property landscapes to guide design [14, 15]. Uncertainty quantification plays a pivotal role here, with techniques like dropout networks providing bounds on predictions, informing risk-aware decision-making [22, 28].

The paradigms also address explainability, where interpretable models reveal feature importance, aiding in the rational design of defects or compositions [10, 20]. Synthesis highlights trade-offs in paradigm selection, such as between exploratory breadth in high-throughput setups and depth in targeted autonomous loops [8, 23].

Uncertainty & interpretability

Uncertainty quantification in materials AI has gained prominence, addressing both aleatoric variability and epistemic limitations through Bayesian frameworks or ensemble methods [16, 22, 28]. In representation learning, this involves propagating uncertainties from data to inferences, ensuring confidence-aware predictions in polycrystalline property modeling or defect engineering [18, 20]. Interpretability complements this by dissecting model decisions, as seen in explainable ML for materials chemistry [10, 26].

The literature synthesizes these elements as essential for trustworthy pipelines, where uncertainty informs interpretability, reducing overconfidence in discovery outcomes [1, 12, 25]. Challenges persist in scaling these to multimodal settings, where coupled uncertainties from diverse sources require integrated handling [4, 6, 32]. Overall, this background reveals an opportunity to elevate uncertainty from a corrective tool to an infrastructural governor of confidence in materials ecosystems.

Proposed conceptual framework

To address the identified gaps, we introduce the Confidence Governance Framework (CGF), an original conceptual architecture that positions uncertainty as a foundational infrastructure for data-driven materials pipelines. The CGF comprises three interconnected layers: the Data Representation Layer, the Inference Steering Layer, and the Discovery Feedback Layer. Each layer embeds uncertainty quantification mechanisms to govern confidence dynamically, ensuring that epistemic risks are not merely quantified but actively shape computational workflows. The infrastructural roles, governance functions, and epistemic outputs associated with each layer of the Confidence Governance Framework are systematically summarized in **Table 1**.

Table 1. Infrastructure Functions of Uncertainty across the Confidence Governance Framework Layers

CGF Layer	Primary Infrastructural Role	Embedded Uncertainty Mechanisms	Goverı Funct

Data Representation Layer	Encoding multimodal materials data into structured AI-readable embeddings	Simulation variance mapping; probabilistic graph weights; multimodal confidence tagging	Calibra represer fidelity; filtratio detec
Inference Steering Layer	Translating representations into predictive and generative outputs	Bayesian inference envelopes; ensemble variance; dropout uncertainty; interpretability scoring	Confid thresh predi routi explora exploi modul
Discovery Feedback Layer	Reintegrating inference outputs into iterative discovery cycles	Adaptive sampling uncertainty maps; experimental deviation metrics; uncertainty propagation loops	Closed recalib data acc steering retrai prioriti
Cross-Layer Governance Channels	Transmitting uncertainty signals across pipeline strata	Layer-to-layer uncertainty propagation; risk aggregation weighting	Glo epist calibra infrastr wide cor harmon

At the core is the Data Representation Layer, which handles the encoding of multimodal materials data into structured forms suitable for AI processing. Here, uncertainty arises from data incompleteness or noise, and the framework conceptualizes it as a modulating factor in representation fidelity. For instance, graph-based embeddings are augmented with uncertainty-aware nodes, where edge weights reflect confidence in atomic interactions derived from simulation variances [3, 5]. This layer facilitates the transition from raw datasets to probabilistic representations, enabling downstream models to inherit calibrated confidences.

The Inference Steering Layer builds upon this by integrating uncertainty into model architectures and decision logics. Rather than treating predictions as point estimates, the CGF employs steering functions that adjust inference paths based on uncertainty thresholds, prioritizing robust outcomes in high-dimensional spaces [16, 22]. This can be conceptualized as a trade-off dynamic between exploration and exploitation, captured in the following symbolic expression:

$$S = argmaxp\,(U\,(p) \cdot I\,(p) - C\,(u)) \quad (1)$$

where S denotes the steered inference path, p p p represents candidate predictions, U(p) is the utility of the prediction, I(p) its interpretability score, and C(u) the cost associated with uncertainty u u u. This expression captures the interaction between predictive value and uncertainty penalties, promoting confidence-governed model behaviors without empirical tuning.

The Discovery Feedback Layer closes the loop by propagating uncertainties back into data acquisition and model refinement. In autonomous systems, this involves adaptive sampling strategies that target high-uncertainty regions, refining pipelines iteratively [13, 19]. Feedback loops are modeled as cyclic dependencies, where output confidences inform input augmentations, fostering resilience against epistemic drifts.

A second formula expresses the feedback dynamics:

$$Ft + 1 = Ft + \Delta D \cdot e - k \cdot ut \quad (2)$$

Here, Ft+1 is the updated framework state at time t+1, Ft the current state, ΔD data increments from feedback, ut prevailing uncertainty, and k a decay constant representing governance strength. This may be expressed as a damped amplification of discoveries, where high uncertainty tempers aggressive updates, ensuring stable pipeline evolution.

The layered embedding of uncertainty across representations, inference steering, and feedback

recalibration is structurally synthesized within the Confidence Governance Framework, as illustrated in **Figure 1**.

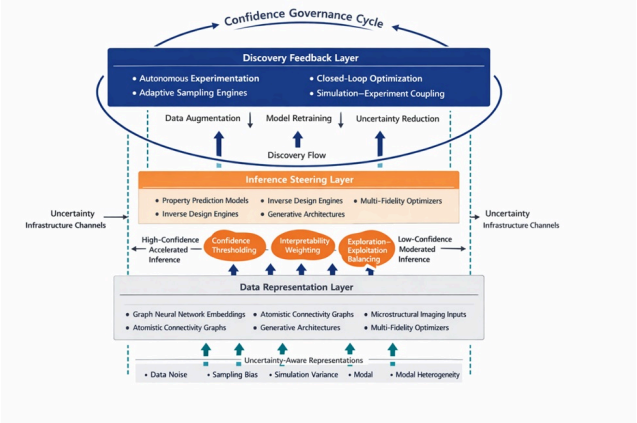

Figure 1. Conceptual Architecture of the Confidence Governance Framework (CGF) for Uncertainty-Aware AI-Driven Materials Discovery

Figure 1 Conceptual systems architecture of the Confidence Governance Framework (CGF), illustrating uncertainty as a governing infrastructure across AI-driven materials discovery pipelines. The framework is organized into three interconnected layers: the Data Representation Layer, where multimodal materials data are encoded into uncertainty-aware embeddings; the Inference Steering Layer, where predictive pathways are modulated by confidence thresholds and interpretability weights; and the Discovery Feedback Layer, where uncertainty propagation informs adaptive sampling and pipeline recalibration. Bidirectional dashed conduits depict uncertainty transmission across layers, while solid arrows indicate discovery flows from data ingestion to experimental steering. The architecture highlights how infrastructural confidence governance regulates epistemic risk, resource allocation, and exploration depth in autonomous and high-throughput materials engineering systems.

A final conceptual formula addresses epistemic risk aggregation across layers:

$$R = \sum lwl \cdot ul \cdot (1 - \gamma l) \quad (3)$$

with R as total risk, l indexing layers, wl weights, ul layer-specific uncertainties, and γl governance factors. This captures the weighted interplay of uncertainties, reduced by infrastructural interventions, underscoring the framework's interpretive emphasis on risk-aware discovery steering.

Analytical Implications

Implications for discovery pipelines

The Confidence Governance Framework (CGF) offers interpretive insights into the dynamics of data-driven discovery pipelines, where uncertainty infrastructure recalibrates the flow from representation to outcome. In high-throughput systems, this implies a shift toward confidence-modulated screening, where candidate materials are prioritized not solely by predicted properties but by aggregated epistemic assurances [7, 9]. Such steering logics can mitigate resource inefficiencies, as low-confidence predictions are deferred or rerouted for additional data augmentation, enhancing overall pipeline throughput without empirical overhead [13, 14].

Computationally, this manifests in optimized inverse design workflows, where uncertainty channels inform the inversion process, balancing explorative breadth with risk-aware convergence [21, 29]. For instance, in autonomous discovery, the framework's feedback layers suggest interpretive trade-offs, such as delaying closed-loop iterations until uncertainty thresholds are met, thereby aligning computational efforts with discovery reliability [19, 31].

Epistemic risk structures

Analytically, the CGF illuminates epistemic risk structures in materials AI, framing uncertainty as a scaffold for risk aggregation across pipeline stages. This perspective reveals how unaddressed uncertainties compound, potentially amplifying biases in multimodal datasets or simulation couplings [22, 24]. By embedding governance factors, the framework interprets risk as a distributed entity, where layer-specific mitigations—such as probabilistic embeddings in representations—dampen propagation effects [5, 16].

This can be expressed as an interpretive aggregation:

$$E = \prod l\,(1 + rl \cdot ul) - 1 \quad (4)$$

where E captures cumulative epistemic exposure, rl the risk multiplier per layer l and ul the uncertainty contribution. This formula highlights the multiplicative interactions in risk buildup, underscoring the need for infrastructural interventions to maintain sub-unitary growth.

Representation-inference interactions

The framework's implications extend to the interplay between representations and inferences, where uncertainty infrastructure governs the fidelity of mappings in graph-based models [3, 6]. Analytically, this suggests enhanced interpretability through confidence-weighted features, allowing for the dissection of model decisions in complex materials like polycrystals or defects [18, 20]. In data-scarce regimes, such interactions imply adaptive learning paradigms that leverage uncertainty to guide transfer from multi-fidelity sources [11, 27].

Infrastructure trade-offs

Finally, the CGF analytically surfaces trade-offs in computational infrastructures, such as the balance between model complexity and uncertainty manageability [1, 12]. High-dimensional representations, while expressive, may inflate epistemic costs unless governed, as seen in explainable ML applications [10, 26]. This interpretive lens promotes resource-efficient designs, where uncertainty as infrastructure optimizes the allocation between data curation, model training, and discovery validation [8, 17].

Results and Discussion

The conceptualization of uncertainty as infrastructure within the Confidence Governance Framework (CGF) advances contemporary discourse in computational materials engineering by repositioning confidence not as an auxiliary statistical descriptor but as a governing systems property embedded across discovery architectures. In doing so, the framework responds directly to the escalating epistemic complexity introduced by high-throughput computation, multimodal data fusion, and autonomous experimentation platforms. As materials informatics ecosystems scale, predictive outputs increasingly travel across extended interpretive chains—from simulation to surrogate inference to generative design—where each translation layer introduces compounding confidence distortions. By structurally integrating uncertainty governance within these translational pathways, CGF reframes reliability as an infrastructural design outcome rather than a retrospective analytical correction.

This repositioning carries significant implications for how opacity is interpreted in AI-guided materials pipelines. High-capacity architectures—particularly graph neural networks, equivariant encoders, and multimodal foundation models—have demonstrated unprecedented predictive performance across compositional and structural domains. However, performance gains often coincide with declining interpretability, producing inference regimes where high confidence scores mask representational fragility. Within closed-loop discovery systems, such opacity propagates downstream, steering experimental prioritization, resource allocation, and design optimization under potentially distorted epistemic premises. CGF intervenes at this juncture by embedding confidence diagnostics directly into representation learning, inference weighting, and feedback steering layers, ensuring that predictive authority remains coupled to epistemic traceability.

Unlike post-hoc uncertainty quantification strategies—which typically append Bayesian intervals, ensemble variance metrics, or dropout approximations after model training—the CGF distributes confidence governance across the pipeline lifecycle. This distribution enables dynamic recalibration as data regimes evolve. For example, when surrogate models extrapolate into sparsely sampled compositional regions, the framework’s governance nodes can attenuate inference weightings, redirect sampling, or trigger density-compensatory acquisition strategies. In this sense, uncertainty becomes a steering signal rather than a passive descriptor, modulating exploration velocity, diversity penetration, and optimization depth across discovery trajectories.

Implementation, however, introduces non-trivial infrastructural challenges. Materials datasets are inherently heterogeneous, spanning density functional theory outputs, experimental microstructures, spectroscopy, and process metadata. Each modality carries distinct noise signatures, resolution limits, and sampling biases. Embedding uniform confidence governance across such heterogeneous substrates necessitates modular epistemic translation layers capable of harmonizing uncertainty semantics. In perovskite screening, for instance, uncertainty may derive from thermodynamic metastability approximations, whereas in catalyst discovery it may emerge from surface reconstruction dynamics or measurement variability. CGF therefore requires context-adaptive governance schemas

that translate domain-specific epistemic uncertainties into interoperable steering metrics.

These heterogeneities further complicate multimodal fusion architectures. As image-derived microstructures merge with graph-encoded atomic networks and text-mined synthesis descriptors, uncertainty propagation becomes non-linear. Confidence distortions in one modality may amplify or dampen inference stability in another, producing emergent epistemic behaviors not reducible to individual data sources. The CGF's layered architecture addresses this through cross-modal confidence arbitration nodes, which evaluate concordance across modalities before propagating inference authority downstream. Such arbitration mechanisms are particularly critical in inverse design pipelines, where generative outputs depend on coherent multimodal embeddings.

Looking forward, the extension of CGF into foundation model ecosystems represents a critical frontier. Pre-trained materials representations—trained on expansive computational corpora—are increasingly deployed as transferable embedding infrastructures for downstream tasks. While these models enhance scalability and reduce training overhead, they also risk encoding latent confidence biases rooted in uneven training distributions. Embedding governance layers within foundation architectures would enable confidence-aware transfer learning, where downstream inferences are weighted not only by predictive alignment but by epistemic compatibility between source and target domains. This would enhance generalizability while mitigating misapplication risks in underrepresented materials classes.

The framework's feedback orientation also invites reconsideration of human-AI epistemic collaboration. Autonomous laboratories and closed-loop optimization platforms increasingly minimize human intervention to accelerate throughput. Yet expert intuition remains uniquely sensitive to contextual anomalies, mechanistic plausibility, and tacit knowledge not encoded in datasets. By integrating hybrid governance nodes—where human oversight modulates machine confidence thresholds—the CGF envisions collaborative steering architectures. In such systems, automated pipelines conduct high-velocity screening while domain experts intervene at epistemic inflection points, recalibrating confidence trajectories and preventing premature convergence.

Broader implications extend into infrastructural sustainability. Computational discovery campaigns often expend substantial resources exploring low-confidence regions without governance modulation, leading to epistemically redundant simulations and energy-intensive computation. By leveraging uncertainty as an exploration regulator, CGF promotes resource-aware discovery, directing computational expenditure toward epistemically fertile zones. This aligns with emerging sustainability imperatives in scientific computing, where energy efficiency and carbon accountability are becoming integral evaluation metrics.

The framework also reorients evaluation paradigms in materials AI. Conventional benchmarking prioritizes predictive accuracy, mean absolute error, or discovery yield. CGF suggests augmenting these metrics with governance robustness indicators—measures of how effectively pipelines maintain epistemic calibration under distributional shift, sparse sampling, or multimodal discordance. Such metrics would incentivize infrastructures that not only discover materials efficiently but do so with traceable confidence integrity.

Finally, the CGF contributes to philosophical discourse on machine-mediated scientific knowledge. By infrastructuralizing uncertainty, the framework challenges deterministic narratives of AI discovery, emphasizing instead that all computational inference operates within governed confidence envelopes. This reframing underscores that scientific validity in AI-accelerated materials engineering is co-produced by data, models, and governance logics—not solely by predictive performance.

Conclusion

In summary, this manuscript has advanced a conceptual reconfiguration of uncertainty within computational materials engineering by positioning it as a governing infrastructure rather than a peripheral statistical artifact. Through the Confidence Governance Framework, uncertainty is architecturally embedded across representation learning, inference generation, and discovery steering layers, forming an integrated confidence ecosystem that shapes how knowledge is produced, interpreted, and operationalized in data-driven materials pipelines.

By articulating a layered governance architecture, the framework provides systems-level interpretive clarity into how confidence propagates, transforms, and accumulates across computational workflows. Representation layers encode epistemic provenance; inference layers translate uncertainty into predictive authority; and feedback layers recalibrate exploration trajectories in response to evolving confidence landscapes. This integrative structuring transforms uncertainty from a passive descriptor into an active infrastructural signal capable of steering discovery dynamics.

The implications for materials engineering ecosystems are substantial. Confidence-governed pipelines promise enhanced interpretability in black-box architectures, improved generalizability across heterogeneous materials domains, and more resilient decision-making in autonomous discovery environments. By coupling predictive outputs with epistemic traceability, the framework strengthens translational bridges between computational forecasts and experimental realization—an essential requirement for deploying AI-generated materials in real-world applications.

Moreover, the CGF introduces a sustainability dimension to computational discovery. Confidence-aware steering reduces epistemically redundant exploration, optimizing computational expenditure and aligning discovery infrastructures with energy-efficient scientific computing practices. As materials AI continues to scale, such governance mechanisms will be critical in balancing discovery acceleration with responsible resource utilization.

Looking forward, the maturation of AI-driven materials science will depend not only on algorithmic sophistication or data expansion but on the governance architectures that regulate epistemic reliability. Frameworks such as CGF offer conceptual blueprints for embedding confidence accountability into next-generation discovery systems, including foundation models, autonomous laboratories, and hybrid human-AI research ecosystems.

As computational and experimental infrastructures converge, the ability to govern uncertainty coherently will determine the credibility, efficiency, and societal impact of accelerated materials innovation. By foregrounding confidence as infrastructure, this work contributes to a broader paradigm shift—one in which uncertainty is no longer treated as a limitation to be minimized, but as an epistemic resource to be structured, interpreted, and strategically leveraged in the pursuit of scientific discovery.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 27 Jan 2022 Revised: 19 Jun 2022 Accepted: 02 Aug 2022
Published online: 18 September 2022

References

Schmidt J, Marques MRG, Botti S, Marques MAL. Recent advances and applications of machine learning in solid-state

materials science. npj Comput Mater. 2019;5:83. https://doi.org/10.1038/s41524-019-0221-0.

Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C. Machine learning in materials informatics: Recent applications and prospects. npj Comput Mater. 2017;3:54. https://doi.org/10.1038/s41524-017-0056-5.

Dai M, Demirel MF, Liang Y, Hu JM. Graph neural networks for an accurate and interpretable prediction of the properties of polycrystalline materials. npj Comput Mater. 2021;7:103. https://doi.org/10.1038/s41524-021-00574-w.

Fung V, Zhang J, Juarez E, Sumpter BG. Benchmarking graph neural networks for materials chemistry. npj Comput Mater. 2021;7:84. https://doi.org/10.1038/s41524-021-00554-0.

Choudhary K, DeCost B. Atomistic line graph neural network for improved materials property predictions. npj Comput Mater. 2021;7:185. https://doi.org/10.1038/s41524-021-00650-1.

Gu GH, Jang J, Noh J, Walsh A, Jung Y. Perovskite synthesizability using graph neural networks. npj Comput Mater. 2022;8:71. https://doi.org/10.1038/s41524-022-00757-z.

Yeo BC, Nam H, Nam H, Kim MC, Lee HW, Kim SC, et al. High-throughput computational-experimental screening protocol for the discovery of bimetallic catalysts. npj Comput Mater. 2021;7:137. https://doi.org/10.1038/s41524-021-00605-6.

Shen ZX, Su C, He L. High-throughput computation and structure prototype analysis for two-dimensional ferromagnetic materials. npj Comput Mater. 2022;8:132. https://doi.org/10.1038/s41524-022-00813-8.

Brunin G, Ricci F, Ha VA, Rignanese GM, Hautier G. Transparent conducting materials discovery using high-throughput computing. npj Comput Mater. 2019;5:63. https://doi.org/10.1038/s41524-019-0200-5.

Zhong X, Gallagher B, Liu S, Kailkhura B, Hiszpanski A, Han TYJ. Explainable machine learning in materials science. npj Comput Mater. 2022;8:204. https://doi.org/10.1038/s41524-022-00884-7.

Pilania G. Machine learning in materials science: From explainable predictions to autonomous design. Comput Mater Sci. 2021;193:110360. https://doi.org/10.1016/j.commatsci.2021.110360.

Mishin Y. Machine-learning interatomic potentials for materials science. Acta Mater. 2021;214:116980. https://doi.org/10.1016/j.actamat.2021.116980.

Ren E, Guilbaud P, Coudert FX. High-throughput computational screening of nanoporous materials in targeted applications. Digit Discov. 2022;1:355-74. https://doi.org/10.1039/d2dd00018k.

Boland TM, Singh AK. Computational synthesis of 2D materials: A high-throughput approach to materials design. Comput Mater Sci. 2022;207:111238. https://doi.org/10.1016/j.commatsci.2022.111238.

Wolloch M, Losi G, Chehaimi O, Yalcin F, Ferrario M, Righi MC. High-throughput generation of potential energy surfaces for solid interfaces. Comput Mater Sci. 2022;207:111302. https://doi.org/10.1016/j.commatsci.2022.111302.

Garcia-Cardona C, Fernandez-Godino MG, O'Malley D, Bhattacharya T. Uncertainty bounds for multivariate machine learning predictions on high-strain brittle fracture. Comput Mater Sci. 2022;201:110883. https://doi.org/10.1016/j.commatsci.2021.110883.

Ong SP. Accelerating materials science with high-throughput computations and machine learning. Comput Mater Sci. 2019;161:143-50. https://doi.org/10.1016/j.commatsci.2019.01.013.

Hestroffer JM, Charpagne MA, Latypov MI, Beyerlein IJ. Graph neural networks for efficient learning of mechanical properties of polycrystals. Comput Mater Sci. 2022;217:111894. https://doi.org/10.1016/j.commatsci.2022.111894.

Omee S, Louis SY, Zhang S, Wei CH, Hu J. Neural network based active learning for high-throughput alloy development. Comput Mater Sci. 2022;204:111140. https://doi.org/10.1016/j.commatsci.2021.111140.

Frey NC, Akinwande D, Jariwala D, Shenoy VB. Machine learning-enabled design of point defects in 2D materials for quantum and nonlinear photonics. Nat Mater. 2021;20:199-204. https://doi.org/10.1038/s41563-020-00840-1.

Saidi P, Quadrelli EA. Machine learning discovery of chemical catalysts for selective catalytic reduction processes. Nat Mach Intell. 2022;4:537-46. https://doi.org/10.1038/s42256-022-00494-2.

Luan X, Qin C, Chang F, Zhang H, Jin X, Tong X. Uncertainty quantification in molecular simulations with dropout neural network potentials. npj Comput Mater. 2020;6:107. https://doi.org/10.1038/s41524-020-00390-8.

Bartel C, Trewartha A, Wang Q, Dunn A, Porter A, Sutton C, et al. Physical descriptor for the Gibbs energy of inorganic crystalline solids and temperature-dependent materials chemistry. Nat Commun. 2020;11:3140. https://doi.org/10.1038/s41467-020-17114-9.

Dunn A, Wang Q, Ganose A, Dagdelen J, Jain A. Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm. npj Comput Mater. 2020;6:138. https://doi.org/10.1038/s41524-020-00406-3.

Goodall RE, Lee AA. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. Nat Commun. 2020;11:6280. https://doi.org/10.1038/s41467-020-19964-7.

Chen D, Müller J, Exertier C, Bagherian A, Cozzolino D, Vidyasagar A, et al. Deep learning aided high-throughput defect detection in ceramic films for solid-state batteries. npj Comput Mater. 2022;8:224. https://doi.org/10.1038/s41524-022-00922-4.

Pilania G, Gubernatis JE, Lookman T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. Comput Mater Sci. 2017;129:156-63. https://doi.org/10.1016/j.commatsci.2016.12.004.

Zhang Y, Ling C. A strategy to apply machine learning to small datasets in materials science. npj Comput Mater. 2018;4:25. https://doi.org/10.1038/s41524-018-0081-8.

Dan Y, Zhao Y, Li X, Li S, Hu M, Hu J. Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse materials design. Nat Commun. 2020;11:2318. https://doi.org/10.1038/s41467-020-16056-y.

Goodall RE, Parackal AS, Faber FA, Armiento R, Lee AA. A scalable representation of local energy landscapes for machine learning. npj Comput Mater. 2022;8:40. https://doi.org/10.1038/s41524-022-00705-x.

Podryabinkin EV, Shapeev AV. Active learning of linearly parametrized interatomic potentials. Comput Mater Sci. 2017;140:171-9. https://doi.org/10.1016/j.commatsci.2017.08.031.

Chen L, Tran H, Batra R, Kim C, Ramprasad R. Machine learning models for the prediction of energy, electron affinity, and ionization energy of molecules. Sci Adv. 2019;5:eaav6494. https://doi.org/10.1126/sciadv.aav6494.