

REVIEW

Open access

Computational and Data-Driven Materials Engineering: Multimodal Materials Datasets, Integration Frameworks, and Discovery Potential

Maria Gonzalez^{1*}, Javier Ruiz¹, Lucia Torres²

Abstract

The field of computational and data-driven materials engineering has transformed from traditional high-throughput simulations to sophisticated ecosystems integrating machine learning with multimodal datasets for accelerated discovery. This review synthesizes recent advancements in materials informatics, emphasizing the role of graph neural networks and deep learning in processing complex structural and property data. We examine multimodal datasets that combine experimental, computational, and textual modalities, enabling robust representation learning and uncertainty quantification. Integration frameworks are discussed, including active learning loops and multi-fidelity models that bridge simulation and experiment, addressing challenges like data sparsity and distribution shifts. The discovery potential is highlighted through applications in property prediction, inverse design, and autonomous systems, such as identifying stable alloys and energy materials. By providing an original synthesis of these elements, this article underscores the shift toward closed-loop workflows that enhance generalizability and interpretability, while identifying gaps in handling finite-temperature stability and disordered systems. Ultimately, these approaches promise to expand the known materials space by orders of magnitude, fostering innovations in sustainable technologies.

Keywords Materials informatics, Uncertainty quantification, Machine learning, Active learning, Graph neural networks, Multimodal datasets

*Correspondence:

Maria Gonzalez
maria.gonzalez@gmail.com

¹ Department of Materials Informatics, Faculty of Engineering, University of Granada, Granada, Spain

² Department of Computational Materials Simulation, Faculty of Engineering, University of Seville, Seville, Spain

Introduction

The discovery and engineering of novel materials underpin nearly every technological transition in human history, from early metallurgical innovations to contemporary advances in semiconductors, quantum devices, and sustainable energy systems. Historically, materials development progressed through empirical trial-and-error experimentation guided by accumulated domain expertise, thermodynamic heuristics, and phenomenological models [1]. While such approaches yielded transformative breakthroughs, they were inherently constrained by low throughput, fragmented data accumulation, and the bounded intuition of human investigators. As materials challenges became increasingly complex—spanning multicomponent alloys, metastable phases, and nanoscale architectures—the limitations of conventional discovery pipelines became more pronounced.

A fundamental barrier arises from the combinatorial vastness of materials space. Estimates suggest that the number of thermodynamically plausible inorganic compounds alone exceeds 10^{100} , rendering exhaustive experimental or

computational exploration infeasible [2, 3]. Even with advances in synthesis and characterization, the cost, time, and resource demands of systematically probing such spaces far exceed practical limits. This combinatorial explosion catalyzed the transition toward computational screening, where first-principles simulations could evaluate candidate materials prior to experimental realization.

Density functional theory (DFT) emerged as the cornerstone of this computational paradigm, enabling quantum-mechanical estimation of formation energies, electronic structures, and mechanical properties across thousands of compounds with reasonable accuracy [4]. High-throughput DFT workflows institutionalized large-scale virtual screening, systematically populating databases of computed materials properties. This shift was formalized through global initiatives such as the Materials Genome Initiative (MGI), which articulated a strategic vision for integrating computation, experiment, and digital infrastructure to accelerate materials discovery by an order of magnitude [5, 6]. The MGI catalyzed the development of interoperable data repositories, standardized workflows, and collaborative platforms that laid the groundwork for modern materials informatics ecosystems.

In the past decade, the field has undergone a second transformation with the emergence of data-driven methodologies—often described as the “fourth paradigm” of scientific discovery, complementing experiment, theory, and simulation [3, 7]. Rather than relying solely on physics-based simulations, these approaches leverage machine learning (ML) to extract structure–property relationships directly from data. By training surrogate models on high-throughput computational outputs, ML systems can approximate DFT-level predictions at a fraction of the computational cost [2, 8]. This capability has enabled rapid estimation of formation energies, band gaps, elastic tensors, and thermal properties across expansive compositional domains [9–11]. Consequently, ML has shifted computational materials science from deterministic simulation toward probabilistic inference and predictive analytics.

The scalability of these models is enabled by large curated repositories such as the Materials Project, the Open Quantum Materials Database (OQMD), and the Automatic FLOW for Materials Discovery (AFLOW) consortium, each aggregating millions of computed entries [4, 12]. These infrastructures serve not merely as data warehouses but as training substrates for representation learning, allowing models to interpolate within known chemical regimes and cautiously extrapolate toward unexplored compounds [13, 14]. However, interpolation-centric learning introduces epistemic risks when models encounter distributional gaps—an issue that has intensified interest in uncertainty quantification and domain generalization.

A major conceptual evolution within this landscape is the rise of multimodal materials datasets, which integrate heterogeneous data sources into unified learning environments. Unlike unimodal datasets restricted to compositional or structural descriptors, multimodal corpora incorporate crystallographic structures, spectroscopic signatures, microstructural imagery, synthesis parameters, and even textual knowledge extracted from scientific literature [15–17]. This integration addresses critical blind spots in traditional descriptor systems. For example, compositional vectors alone cannot distinguish polymorphic phases, whereas structural graphs derived from crystallographic information files (CIFs) encode geometric and topological distinctions [8, 18].

Graph neural networks (GNNs) have become the dominant architectures for exploiting such structural data. By representing materials as atomic graphs—where nodes encode elements and edges encode interatomic interactions—GNNs capture local coordination environments and long-range periodicity [18, 19]. Advanced variants incorporate rotational equivariance, angular message passing, and physics-informed constraints, ensuring symmetry preservation and improved generalization. Benchmark studies across datasets such as QM9 and Materials Project demonstrate that graph-based models consistently outperform traditional hand-crafted descriptors in property prediction tasks [9, 13]. These advances signal a shift from feature engineering toward automated representation discovery.

Operationalizing multimodal intelligence requires integrative computational frameworks that couple data ingestion, model inference, and experimental validation. Hybrid simulation–ML pipelines exemplify this integration. For instance, Gaussian process regression (GPR) can refine DFT predictions while quantifying predictive uncertainty, enabling adaptive

screening strategies [20, 21]. Active learning frameworks iteratively identify high-value data points for simulation or experiment, optimizing resource allocation across discovery cycles [20, 21]. Multi-fidelity modeling further enhances efficiency by fusing low-fidelity approximations (e.g., empirical potentials) with high-accuracy quantum calculations through co-kriging or hierarchical learning schemes [2, 22].

Transfer learning introduces an additional scaling mechanism. Models pretrained on large computational datasets—such as QM9—can be fine-tuned on scarce experimental measurements, improving predictive robustness in data-limited regimes [10, 22]. Such strategies are particularly critical in materials science, where experimental datasets remain orders of magnitude smaller than those in computer vision or natural language processing [7, 10, 22]. Incorporating approximate physical estimates, such as generalized gradient approximation (GGA) band gaps, as auxiliary features further enhances predictive performance without increasing computational burden [10].

The downstream discovery implications of these paradigms are substantial. Machine learning-driven screening has identified hundreds of thousands of thermodynamically stable materials from millions of hypothetical candidates, dramatically expanding the convex hull of known compounds [13]. In structural alloys, ML-guided design has enabled the identification of high-strength, ductile titanium systems with optimized performance trade-offs [23]. Energy materials research has similarly benefited, with data-mining frameworks accelerating the discovery of battery electrodes, catalysts, and thermoelectric compounds [24]. Autonomous laboratories extend this paradigm by embedding ML within closed-loop experimentation systems that dynamically balance exploration and exploitation [15, 20].

Reliability within such systems depends critically on uncertainty quantification. Techniques including ensemble variance, Bayesian inference, and Gaussian process uncertainty estimates enable models to signal confidence levels and detect out-of-distribution inputs [11, 15]. These tools are essential for managing distribution shifts as databases evolve and new chemistries are incorporated. Complementary visualization methods—such as uniform manifold approximation and projection (UMAP)—provide geometric insights into dataset topology, revealing domain gaps and clustering biases [14].

Despite rapid progress, several structural challenges persist. Interpretability remains central to scientific trust, motivating the integration of explainable AI (XAI) techniques such as saliency mapping, attention attribution, and surrogate modeling [21, 25]. Data scarcity continues to constrain model generalizability, driving interest in physics-informed machine learning that embeds conservation laws and thermodynamic constraints within neural architectures [16, 22]. Furthermore, while predictive modeling has matured, true inverse design—where desired properties generate candidate structures—remains emergent, relying on generative frameworks such as variational autoencoders (VAEs) and generative adversarial networks (GANs) [8, 19].

Positioned within this rapidly evolving landscape, the present review advances an integrative synthesis of computational and data-driven materials engineering ecosystems. Rather than focusing on isolated algorithmic advances, we conceptualize the field as an interconnected infrastructure spanning multimodal data substrates, representation learning architectures, integrative simulation frameworks, and autonomous discovery systems. By structuring these components within closed-loop innovation pipelines, this article offers a systems-level perspective that extends beyond prior technique-centric reviews [2, 5, 8]. Drawing on cross-study analyses from recent literature (2017–2023), we highlight emergent synergies—particularly the coupling of representation learning with adaptive experimentation—as foundational drivers of next-generation materials discovery [12, 19, 20].

Collectively, these developments position computational and data-driven materials engineering as a unified discovery ecosystem—one capable of navigating vast chemical design spaces, orchestrating multimodal knowledge, and accelerating innovation across energy, structural, and functional materials domains.

Landscape of Computational & Data-Driven Materials Engineering

Evolution of materials informatics ecosystems

Materials informatics has undergone a profound structural evolution, transitioning from fragmented computational efforts to fully integrated discovery ecosystems powered by machine learning and digital infrastructures [5, 6]. Early computational materials initiatives were primarily oriented around high-throughput density functional theory (DFT) calculations designed to populate searchable repositories of materials properties. Foundational platforms such as the Materials Project and the Open Quantum Materials Database (OQMD) exemplified this first wave, systematically generating formation energies, electronic structures, and thermodynamic stability metrics across thousands of inorganic compounds [4, 12]. These repositories established the empirical backbone of modern materials informatics by enabling large-scale property screening and comparative thermodynamic analyses.

Crucially, these infrastructures were developed in alignment with FAIR data principles—ensuring that materials datasets were findable, accessible, interoperable, and reusable. Such standardization enabled the transition from static databases to dynamic knowledge systems capable of supporting data mining, surrogate modeling, and predictive analytics [2, 9]. Surrogate models trained on these datasets demonstrated the ability to reproduce quantum-mechanical predictions with near-chemical accuracy, dramatically reducing computational costs associated with conventional simulations.

Recent advancements have extended these repositories into fully operational ecosystems in which machine learning serves as the connective substrate linking data generation, predictive modeling, experimental validation, and iterative optimization [5, 26]. In this ecosystem paradigm, computational screening is no longer an isolated endpoint but a component within closed-loop discovery architectures. Emerging intelligence systems now incorporate natural language processing (NLP) pipelines to extract synthesis routes, processing conditions, and performance outcomes directly from the scientific literature [25, 27]. These text-mined insights augment structured databases, enabling richer multimodal representations of materials knowledge.

This transition is partly driven by scale asymmetry. Traditional DFT infrastructures, even under high-throughput regimes, are constrained to evaluating on the order of 10^5 compounds annually due to computational expense [4, 13]. In contrast, machine learning surrogates can screen hypothetical materials spaces exceeding 10^9 candidates within comparable timeframes. This exponential scaling advantage has repositioned ML not merely as an efficiency tool but as a fundamental enabler of large-scale exploratory science.

At the foundation of these ecosystems lies the recognition of data as the central discovery asset. Contemporary repositories such as JARVIS and SuperCon extend beyond structural and thermodynamic descriptors to include spectral measurements, synthesis metadata, and textual annotations [7, 8]. These multimodal inputs facilitate cross-domain inference and enhance representation richness. However, they also introduce epistemic challenges. Dataset biases—such as the overrepresentation of thermodynamically stable phases or publication-positive results—can distort predictive models and inflate apparent performance [7, 27].

To mitigate such risks, advanced validation methodologies have been introduced. Techniques including asymmetric validation embedding and leave-one-cluster-out cross-validation evaluate model generalizability across chemically distinct subspaces rather than random splits [7, 14]. These approaches ensure that trained systems retain predictive reliability when extrapolating into sparsely sampled or novel compositional domains, reinforcing trust in large-scale screening outputs.

Key machine learning techniques and representations

Machine learning techniques within materials engineering span a broad methodological spectrum, encompassing supervised regression, classification frameworks, and generative modeling architectures designed for inverse discovery [3, 8]. Among these, deep learning has emerged as the dominant paradigm, particularly in contexts where materials structures can be encoded as graph-based relational systems.

Graph neural networks (GNNs) have demonstrated exceptional performance in modeling crystalline solids due to their ability to represent atoms as nodes and interatomic interactions as edges [18, 19]. Architectures such as Crystal Graph Convolutional Neural Networks (CGCNN) and Atomistic Line Graph Neural Networks (ALIGNN) extend this paradigm by incorporating bond distances, angular relationships, and higher-order geometric interactions. These models have achieved mean absolute errors as low as ~ 0.022 eV/atom in formation energy prediction tasks, approaching the intrinsic uncertainty limits of DFT calculations themselves [9, 13].

Equivariant neural networks further advance this representation paradigm by embedding physical symmetry constraints directly into model architectures. Frameworks such as NequIP enforce rotational and translational equivariance, enabling accurate force-field predictions suitable for molecular dynamics simulations. These models have demonstrated computational accelerations approaching three orders of magnitude relative to ab initio methods, fundamentally altering the feasibility of large-scale atomistic simulations [8, 19].

At the core of these advances lies representation learning. Materials descriptors now span hierarchical abstraction levels—from coarse elemental property vectors to sub-Ångström encodings of atomic environments [6]. Descriptor families such as many-body tensor representations (MBTR) and smooth overlap of atomic positions (SOAP) provide rotationally invariant, information-dense encodings that consistently outperform traditional hand-engineered features across benchmarking tasks [12, 17].

For multimodal datasets, representation fusion strategies play a critical role. Feature concatenation frameworks integrate compositional descriptors with spectral embeddings, microstructural imaging features, or thermodynamic variables to produce unified predictive models [15, 16]. In data-limited environments, augmentation strategies such as crude estimation of properties (CEP) embed approximate physical estimates as auxiliary inputs, reducing predictive error margins by as much as 50% [10, 22].

Interpretability remains a central research priority. Explainable AI (XAI) techniques—including Gradient-weighted Class Activation Mapping (Grad-CAM) and SHapley Additive exPlanations (SHAP)—enable post hoc interrogation of model decision pathways [21, 25]. These tools reveal physically meaningful correlations, such as the microstructural determinants of compressive strength in cementitious systems or electronic orbital contributions in transition metal oxides [16, 21].

Complementing interpretability, uncertainty quantification frameworks provide probabilistic confidence estimates for model outputs. Gaussian process regression variance metrics, ensemble disagreement measures, and Bayesian neural approximations are widely deployed to detect out-of-distribution inputs and guide adaptive data acquisition [14, 20]. Such reliability scaffolds are essential for deploying ML systems in safety-critical engineering contexts.

Multimodal datasets and their role

Multimodal datasets represent one of the most transformative developments in contemporary materials informatics, enabling integrated modeling of structure, processing, performance, and environmental response [7, 28]. By combining heterogeneous data streams, these datasets provide a holistic representation of materials behavior unattainable through unimodal descriptors alone.

Prominent examples include the High-Throughput Experimental Materials (HTEM) database, which integrates optical spectroscopy, electronic measurements, and microstructural imaging, and XASDb, which curates X-ray absorption spectra across diverse material classes [4, 8]. These repositories enable predictive workflows spanning phase identification from diffraction patterns to defect detection in scanning electron microscopy (SEM) imagery [8, 15].

In porous materials research, particularly metal–organic frameworks (MOFs), multimodal integration has enabled genomics-like discovery paradigms. Structural topologies, adsorption properties, and text-mined synthesis parameters

are fused to map performance landscapes across vast design spaces [4]. Such integrative modeling accelerates the identification of high-capacity gas storage and separation materials.

Despite these advances, data scarcity remains a structural bottleneck. Certain subfields—such as concrete materials science—report median dataset sizes of approximately 174 samples, limiting deep learning scalability [16]. To address this constraint, natural language processing pipelines such as ChemDataExtractor mine experimental data from published literature, augmenting structured datasets with extracted synthesis and performance metrics [8, 27]. Generative models, including generative adversarial networks (GANs), further expand datasets through synthetic data generation. **Table 1** summarizes the main data modalities used in modern materials informatics, highlighting the typical learning tasks they support and the failure modes they introduce when used in isolation.

Table 1. Multimodal materials data modalities and what they enable in discovery workflows

Modality (input)	Common representations	Typical ML tasks enabled	Primary value	Common pitfalls / bias risks
Crystal structures (CIF)	atomic graphs; invariant descriptors	formation energy, band gap, stability screening	phase-sensitive structure–property mapping	limited finite-T/disorder realism; database drift [7, 14]
Spectra (XRD/XAS/Raman)	1D signal embeddings	phase ID, local environment inference, oxidation state hints	bridges experiment ↔ model space	instrument variability; label noise; domain shift [4, 15]
Microstructure images (SEM/TEM/optical)	CNN features; segmentation masks	defect detection, microstructure–property links	captures processing history signatures	dataset scarcity; annotation burden; confounding [16, 21]
Process metadata	tabular descriptors	process–property prediction, synthesis optimization	injects manufacturability constraints	missingness; nonstandard reporting; selection bias [27]
Text (papers/patents)	NLP embeddings; extracted entities	synthesis route mining; knowledge graphs	scales evidence extraction beyond curated DBs	publication bias; inconsistent nomenclature [25, 27]

Large-scale multimaterial datasets such as DFT-10B for binary alloys demonstrate the power of integrated data infrastructures, achieving cross-chemical prediction errors below 3 meV/atom [12]. However, these datasets also expose systemic biases, particularly mission-driven computation focused on thermodynamically stable or technologically relevant compounds [7, 14]. Bias auditing and dataset balancing thus remain essential for ensuring equitable exploration of materials space.

Applications in property prediction and screening

Machine learning applications in materials engineering span predictive analytics, design optimization, and high-throughput screening workflows [10, 11]. Property prediction remains the most mature domain, with models estimating thermal, mechanical, and electronic properties across diverse material classes.

In alloy systems, ML frameworks predict lattice thermal conductivity with average factor deviations as low as 1.38, outperforming traditional empirical models [10]. Civil engineering applications demonstrate similar advances, with regression architectures forecasting compressive strength in concrete using multimodal inputs such as hyperspectral imaging and compositional descriptors [16].

Energy materials represent a particularly high-impact application domain. Machine learning has enabled the discovery of high-strength titanium alloys, superionic conductors, and next-generation electrode materials through accelerated screening pipelines [23, 24]. These efforts highlight ML's capacity to navigate trade-offs between stability, conductivity, and manufacturability.

Screening infrastructures increasingly rely on surrogate modeling to triage vast candidate libraries. The GNoME discovery engine, for instance, identified 381,000 stable structures from over two million hypothetical compounds, dramatically expanding the convex hull of known stable materials [13]. Active learning further enhances screening efficiency by dynamically selecting high-value candidates for simulation or synthesis, balancing exploratory uncertainty reduction with exploitative property optimization [20, 21].

Multi-fidelity modeling strategies optimize computational resource allocation by co-kriging low-fidelity approximations—such as Perdew–Burke–Ernzerhof (PBE) band gaps—with high-accuracy GW calculations [2, 22]. Nevertheless, generalizability remains a central concern. Distributional shifts between training and deployment datasets can degrade predictive accuracy when models encounter novel chemistries [14]. Domain-of-applicability frameworks address this risk by identifying reliable prediction subspaces, reducing error margins by up to twofold [11].

Challenges and best practices

Despite transformative progress, materials informatics ecosystems face persistent structural and methodological challenges. Data imbalance and sparsity frequently induce overfitting, particularly in small experimental datasets [16, 22]. Transfer learning and physics-guided machine learning offer mitigation pathways by embedding domain constraints within model architectures.

Best-practice guidelines increasingly emphasize model parsimony, uncertainty reporting, and rigorous validation. Occam's razor principles advocate for selecting minimally complex models capable of achieving target performance, reducing interpretability barriers and overfitting risk [7, 16]. Hybrid data integration—combining laboratory, field, and simulation data—further enhances model robustness.

In manufacturing domains such as additive manufacturing, machine learning supports process optimization, defect prediction, and parameter tuning. However, deployment requires stringent validation under real-world variability conditions [29]. Collectively, these considerations underscore the field's transition toward integrated, interpretable, and sustainability-oriented discovery systems that align computational innovation with engineering reliability [26, 27].

Autonomous & closed-loop discovery systems

Autonomous discovery systems represent the pinnacle of data-driven materials engineering, embedding ML within closed-loop workflows that iteratively refine hypotheses, acquire data, and validate predictions [5, 20]. These systems automate the scientific method, integrating simulation, experiment, and analysis to minimize human intervention [15, 21]. Central to this is active learning, which selects data points to maximize information gain, balancing exploration (high uncertainty regions) and exploitation (promising property optima) [12, 13].

A conceptual formula for active learning utility, such as Expected Improvement (EI), formalizes this: $EI(x) = \int [\max(0, f(x) - f_{\text{best}})] p(f(x)|D) df$, where p is the posterior probability from a surrogate model (e.g., GPR), f_{best} is the current best value, and D is existing data [20]. This guides targeted design, as in optimizing piezoelectric electrostrains, where EI outperforms random sampling by identifying superior compounds in fewer iterations [15, 20].

Integration frameworks enable these loops by fusing multimodal data streams. For instance, simulation-experiment integration uses ML surrogates to pre-screen candidates before experimental validation [15, 17]. In autonomous laboratories, robotic synthesis and characterization (e.g., for perovskites) feed real-time data to ML models for adaptive refinement [8, 28]. Multi-fidelity approaches hierarchically combine low-cost computations (e.g., empirical potentials) with high-fidelity DFT, using co-kriging to propagate uncertainties [2, 22]. Graph-based frameworks like ALIGNN facilitate this by encoding structures for rapid predictions, supporting zero-shot tasks like ionic conductivity in unseen compositions [9, 13].

The discovery potential is exemplified in closed-loop systems for inverse design. Generative models, such as VAEs or GANs, produce structures from latent spaces conditioned on target properties [8, 19]. In polymers, GANs generate microstructures for optimized optical absorption, achieving 17% improvements [28]. For alloys, active learning with GNNs predicts magnetostriction or abnormal grain growth, capturing interactions across scales [17, 28]. Uncertainty quantification ensures reliability; ensemble methods or trust scores flag out-of-distribution samples, as in formation energy predictions where disagreement illuminates extrapolation limits [14, 21]. **Figure 1** summarizes the end-to-end computational and data-driven materials engineering ecosystem, linking multimodal data substrates to representation learning, integration frameworks, and closed-loop discovery outcomes with uncertainty and governance feedback.

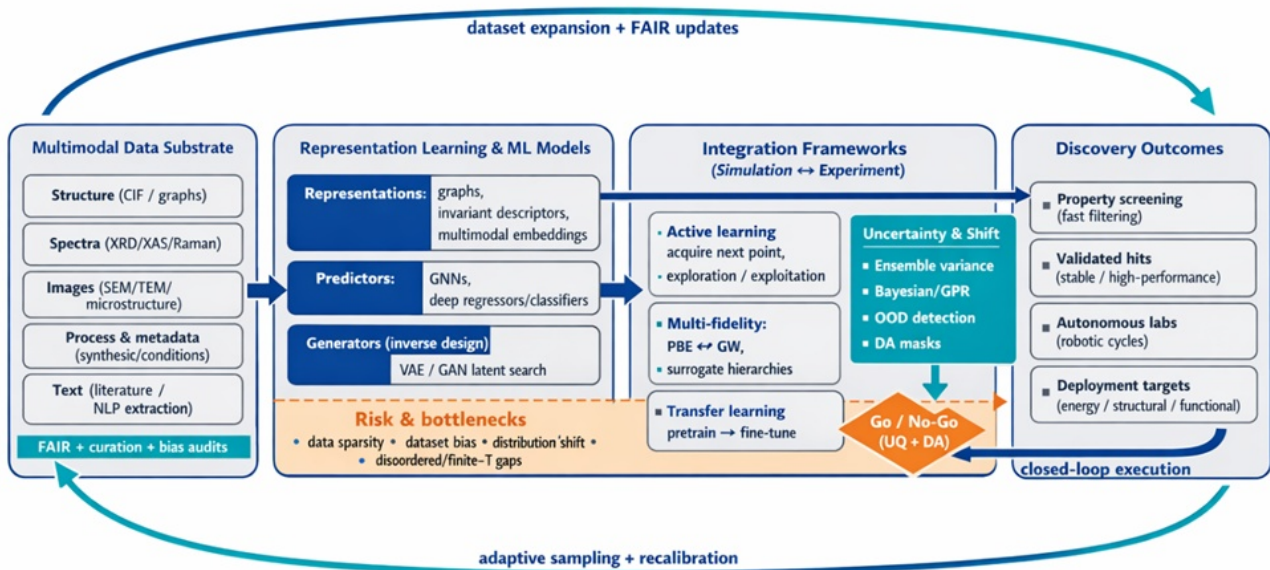


Figure 1. End-to-end computational and data-driven materials engineering ecosystem

Schematic overview of how multimodal materials datasets (structures, spectra, images, processing metadata, and literature-derived text) feed representation learning and graph-based models to produce property predictors, uncertainty estimates, and generative inverse design candidates. Integration frameworks (active learning, multi-fidelity modeling, transfer learning, and domain-of-applicability controls) couple simulation and experiment within closed-loop workflows. The ecosystem converges on validated discovery outcomes (stable materials, energy/structural candidates, and optimized processes) while continuously updating datasets and models through feedback from experiments, simulations, and bias/shift monitoring.

Table 2. Integration frameworks for closed-loop computational–experimental materials discovery

Framework	What it connects	Typical algorithmic choices	What it improves	Key risks / failure modes	Best-practice validation hooks
Active learning	model ↔ new data acquisition	EI/UCB; uncertainty sampling; committee disagreement	sample efficiency; faster convergence [20, 21]	exploitation traps; biased sampling; UQ miscalibration	leave-cluster-out splits; drift monitoring [7, 14]
Multi-fidelity modeling	low-cost ↔ high-accuracy data	co-kriging; hierarchical surrogates	compute efficiency; calibrated trade-offs [2, 22]	fidelity mismatch; inconsistent labels	cross-fidelity consistency checks; error decomposition
Transfer learning	large corpora ↔ scarce targets	pretrain → fine-tune; feature reuse	small-data robustness [10, 22]	negative transfer; hidden confounding	domain similarity audits; ablation studies
Domain of applicability	training domain ↔ deployment domain	DA masks; embedding distance; OOD detection	safer extrapolation [11]	false security; DA boundary drift	periodic re-validation; held-out chemistry clusters
Physics-guided ML	data ↔ constraints	equivariance; conservation priors	stability, plausibility [19, 22]	over-constraint; reduced flexibility	constraint-violation reporting; stress tests
Autonomous labs	prediction ↔ synthesis/characterization	robotic loops; online updating	end-to-end acceleration [15, 28]	noisy measurements; latency; error propagation	uncertainty gating; replicate controls; failure logging

Challenges include handling disordered systems or finite-temperature effects, where ML potentials like NequIP enable molecular dynamics for phase transitions [12, 19]. In concrete science, autonomous systems optimize mixtures by fusing NDT and imaging data [16]. Overall, these systems democratize discovery, expanding stable materials by orders of magnitude while fostering interpretability through XAI [21, 25].

Results and Discussion

Challenges and limitations

While computational and data-driven materials engineering has achieved remarkable strides, several challenges impede its full realization, particularly in handling real-world complexities and ensuring practical applicability [2, 5, 7, 16, 22]. One primary limitation is data scarcity and quality issues, inherent to materials science where datasets are typically small and heterogeneous compared to domains like computer vision [7, 10, 22]. For instance, median dataset sizes in concrete science hover around 174 entries, leading to overfitting and poor generalization [16]. This sparsity is exacerbated by biases in databases, such as overrepresentation of stable, ordered phases or positive results from mission-driven computations, which skew models toward narrow chemical spaces [7, 14, 27]. Distribution shifts between training sets and deployment scenarios further degrade performance; models trained on early Materials Project snapshots fail on later additions due to evolving DFT methodologies or unexplored compositions [14]. To mitigate, researchers advocate for

domain-of-applicability assessments, like those using UMAP embeddings to visualize and constrain prediction subspaces, reducing errors by identifying reliable regions [11].

Another challenge lies in representation learning for complex systems, especially disordered or dynamic materials [12, 17, 19]. Standard GNNs excel in crystalline structures but struggle with amorphous phases, defects, or finite-temperature effects, where atomic environments fluctuate [8, 19]. For example, predicting properties in glasses or polymers requires capturing long-range correlations, often necessitating advanced descriptors like 3-point correlation functions for microstructures [17]. Multimodal integration amplifies this, as fusing disparate data types (e.g., spectral and textual) introduces noise and alignment issues [4, 15]. In small-data contexts, crude approximations help, but they assume underlying correlations that may not hold across modalities [10, 22]. Explainability compounds these issues; while XAI methods like SHAP provide insights, they often highlight spurious correlations in sparse regimes, undermining trust [21, 25].

Uncertainty quantification remains underdeveloped for high-stakes applications, where overconfident predictions can lead to experimental failures [14, 20, 21]. Techniques like ensemble variance work well in controlled settings but falter under distribution shifts, as seen in cross-database validations where errors spike [14]. Active learning loops, while promising, are computationally intensive, requiring careful utility functions to avoid inefficient sampling [20]. In autonomous systems, integration of simulation and experiment introduces latency and error propagation; robotic workflows may generate noisy data, necessitating robust filtering [15, 28]. Moreover, scalability is a bottleneck—deep learning models like those screening 2.2 million structures demand immense resources, limiting accessibility for smaller labs [13, 29].

From a systems perspective, interoperability across ecosystems poses hurdles [5, 26, 27]. Fragmented databases and tools hinder seamless workflows; for example, differing formats between AFLOW and JARVIS complicate multi-fidelity modeling [4, 12]. Big-data approaches in porous materials reveal scalability in storage and querying, but computational overhead for ML training on billions of entries is prohibitive [4]. In additive manufacturing, ML models for process optimization face challenges from variability in machine parameters and materials feedstock, requiring hybrid physics-ML to enforce constraints [29]. Best practices emphasize physics-guided designs, such as incorporating symmetry or conservation laws to regularize models, yet these add complexity without guaranteed improvements [16, 22].

Ethical and practical limitations also warrant discussion. Overreliance on ML may overlook serendipitous discoveries from human intuition, while black-box models risk perpetuating biases in materials selection for applications like energy or infrastructure [7, 25, 27]. In concrete science, challenges include handling lab-field data discrepancies, where environmental factors introduce unmodeled variances [16]. Overall, these limitations highlight the need for balanced ecosystems that prioritize robustness, interpretability, and inclusivity [5, 26].

Future research directions

To overcome current limitations, future efforts in computational and data-driven materials engineering should prioritize scalable, interpretable, and integrated systems [2, 5, 7, 8]. A key direction is advancing multimodal datasets through automated curation and synthesis [4, 15, 27]. Leveraging NLP for extracting unstructured literature data could expand repositories by incorporating synthesis routes and failure modes, addressing sparsity [25, 27]. Synthetic data generation via physics-informed GANs or diffusion models offers promise for augmenting small datasets, particularly in disordered systems where real data is scarce [8, 19, 28]. Standardizing multimodal formats, akin to CIF for structures, would facilitate fusion across modalities, enabling end-to-end workflows from raw spectra to property predictions [15, 17].

Enhancing representation learning for complex materials is another frontier [12, 18, 19]. Developing equivariant GNNs that handle dynamics, such as through time-dependent graphs or tensorial embeddings, could model phase transitions and defects accurately [17, 19]. Multi-scale integrations, linking atomistic (DFT) with mesoscale (phase-field) via ML surrogates, would bridge gaps in applications like additive manufacturing [17, 29]. For inverse design, generative models

conditioned on multiple constraints (e.g., stability and manufacturability) could yield practical candidates, extending beyond current property-focused approaches [8, 13, 24].

Active learning and uncertainty quantification should evolve toward adaptive, multi-objective frameworks [14, 20, 21]. Incorporating human-in-the-loop elements, where experts refine utility functions, could optimize exploration in sparse spaces [15, 20]. Bayesian optimization with multi-fidelity surrogates promises efficiency gains, co-optimizing computational and experimental costs [2, 22]. In autonomous laboratories, real-time ML updates via edge computing could reduce latency, supporting high-throughput campaigns for energy materials [23, 28].

Explainability and generalizability demand focused research [16, 21, 25]. Hybrid XAI-physics models, embedding causal reasoning, could reveal mechanistic insights, as in dissecting microstructural contributions to strength [16, 17]. Benchmarking protocols for distribution shifts, including adversarial testing, would standardize evaluations [11, 14]. Data-centric designs, emphasizing quality over quantity, align with this, advocating for curated subsets that maximize information density [22, 26].

Broader ecosystem developments include federated learning for collaborative databases, preserving proprietary data while enabling collective advancements [5, 27]. Applications in sustainable materials, like recycling-optimized alloys or bio-inspired composites, could leverage these for societal impact [23, 24]. Ultimately, interdisciplinary integrations with robotics and IoT will foster closed-loop ecosystems, accelerating discovery cycles from months to days [15, 28].

Conclusion

Computational and data-driven materials engineering has matured into a transformative ecosystem, harnessing multimodal datasets and ML frameworks to unlock unprecedented discovery potential. From GNN-enabled property predictions to active learning-driven autonomous systems, these approaches have expanded the explorable materials space, yielding innovations in alloys, energy systems, and beyond. Integration frameworks bridge simulation and experiment, while uncertainty tools ensure reliability amid data challenges. Despite limitations in scalability, interpretability, and handling complexity, the field's trajectory toward closed-loop, interpretable workflows promises to redefine materials innovation. By synthesizing these elements, this review underscores the shift from isolated computations to unified ecosystems, paving the way for sustainable technological advancements.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Published online: 18 September 2023

Rights and permissions

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

References

- Liu Y, Zhao T, Ju W, Shi S. Materials discovery and design using machine learning. *J Mater.* 2017;3(3):159-77.
- 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).
- Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. *Nature.* 2018;559(7715):547-55.
<https://doi.org/10.1038/s41586-018-0337-2>.
- Jablonka KM, Ongari D, Moosavi SM, Smit B. Big-Data science in porous materials: Materials genomics and machine learning. *Chem Rev.* 2020;120(16):8066-129.
- Batra R, Song L, Ramprasad R. Emerging materials intelligence ecosystems propelled by machine learning. *Nat Rev Mater.* 2021;6:655-78.
- Ramprasad R, Batra R, Pilania G, Mannodi-Kanakthodi A, Kim C. Machine learning in materials informatics: Recent applications and prospects. *npj Comput Mater.* 2017;3(54).
- Fujinuma N, DeCost B, Hattrick-Simpers J, Lofland SE. Why big data and compute are not necessarily the path to big materials science. *Commun Mater.* 2022;3(57).
- Choudhury S, Bhethanabotla VR, Senftle TP. Recent advances and applications of deep learning methods in materials science. *npj Comput Mater.* 2022;8(59).
- Griesemer SD, Xia Y, Wolverton C. Accelerating the prediction of stable materials with machine learning. *Nat Comput Sci.* 2023;3:934-45.
- Zhang Y, Ling C. A strategy to apply machine learning to small datasets in materials science. *npj Comput Mater.* 2018;4(25).
- Sutton C, Boley M, Ghiringhelli LM, Rupp M, Vreeken J, Scheffler M. Identifying domains of applicability of machine learning models for materials science. *Nat Commun.* 2020;11(3060).
- Nyshadham C, Rupp M, Bekker B, Shapeev AV, Mueller T, Rosenbrock CW, et al. Machine-learned multi-system surrogate models for materials prediction. *npj Comput Mater.* 2019;5(51).
- Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. *Nature.* 2023;624(7990):80-5.

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(62).

Umehara M, Stein HS, Guevarra D, Newhouse PF, Boyd DA, Gregoire JM. Analyzing machine learning models to accelerate generation of fundamental materials insights. *npj Comput Mater.* 2019;5(34).

Li Z, Yoon J, Zhang R, Rajabipour F, Srubar WV III, Dabo I, et al. Machine learning in concrete science: Applications, challenges, and best practices. *npj Comput Mater.* 2022;8(127).

Cheng S, Jiao Y, Ren Y. Data-driven learning of 3-point correlation functions as microstructure representations. *Acta Mater.* 2022;229(117800).

Lu S, Zhou Q, Guo Y, Zhang Y, Wu Y, Wang J. Graph neural networks in learning materials science. *npj Comput Mater.* 2020;6(186).

Reiser P, Neubert M, Eberhard A, Torresi Z, Zhou C, Shao C, et al. Graph neural networks for materials science and chemistry. *Commun Mater.* 2022;3(93).

Lookman T, Balachandran PV, Xue D, Yuan R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Comput Mater.* 2019;5(21).

Kailkhura B, Gallagher B, Kim S, Hiszpanski A, Han TY-J. Reliable and explainable machine-learning methods for accelerated material discovery. *npj Comput Mater.* 2019;5(108).

Chen Y, Hong T, Gopal CB, Han TY-J. Small data machine learning in materials science. *npj Comput Mater.* 2023;9(42).

Liu Y, Esan OC, Pan Z, An L. Machine learning for advanced energy materials. *Energy AI.* 2021;3(100049).

Zou C, Li J, Wang WY, Zhang Y, Lin D, Yuan R, et al. Integrating data mining and machine learning to discover high-strength ductile titanium alloys. *Acta Mater.* 2021;202:211-21.

Zhong X, Gallagher B, Liu S, Kailkhura B, Hiszpanski A, Han TY-J. Explainable machine learning in materials science. *npj Comput Mater.* 2022;8(204).

Chen W, Iyer A, Bostanabad R. Data centric design: A new approach to design of microstructural material systems. *engineering.* 2022;10:89-98.

Klenam DEP, Asumadu TK, Vandadi M, Rahbar N, McBagonluri F, Soboyejo WO. Data science and material informatics in physical metallurgy and material science: An overview of milestones and limitations. *Results Mater.* 2023;19(100455).

Deagen ME, Walsh DJ, Audus DJ, Kroenlein K, de Pablo JJ, Aou K, et al. Networks and interfaces as catalysts for polymer materials innovation. *Cell Rep Phys Sci.* 2022;3(101126).

Johnson NS, Vulimiri PS, To AC, Zhang X, Brice CA, Kappes BB, et al. Invited review: Machine learning for materials developments in metals additive manufacturing. *AM.* 2020;36(101641).