REVIEW Open access

Computational and Data-Driven Materials Engineering: High-Throughput Computational Screening Platforms, Workflows, and Discovery Outcomes

Hiroshi Nakamura[1*], Yuta Kato[1]

Abstract

The field of computational and data-driven materials engineering has undergone rapid evolution, driven by advancements in high-throughput computational screening, machine learning algorithms, and integrated workflows that accelerate materials discovery. This review synthesizes recent developments in materials informatics, focusing on platforms that enable efficient exploration of vast chemical spaces through automated computations and data analytics. Key areas include the application of graph neural networks and representation learning for property prediction, active learning strategies to optimize experimental feedback loops, and the integration of multimodal datasets for enhanced model accuracy. High-throughput methods have facilitated discoveries in diverse domains, such as superconductors, battery materials, and high-entropy alloys, by combining density functional theory simulations with machine learning surrogates. Autonomous laboratories and closed-loop systems represent a paradigm shift, allowing self-driving experiments that minimize human intervention while maximizing discovery efficiency. Uncertainty quantification plays a critical role in guiding these processes, ensuring reliable predictions amid sparse data. This narrative review structures the landscape into computational ecosystems, workflow integrations, and discovery outcomes, highlighting cross-study synergies. It positions the field at the cusp of scalable, inverse design paradigms, where data-driven insights bridge simulation and experimentation to address grand challenges in materials science.

Keywords Materials informatics, Uncertainty quantification, Machine learning, Active learning, Closed-loop discovery, High-throughput computation

*Correspondence:
Hiroshi Nakamura
hiroshi.nakamura@outlook.com
[1] Department of Computational Materials Engineering, Faculty of Engineering, Nagoya University, Nagoya, Japan

Introduction

Materials engineering has historically progressed through empiricism—an iterative paradigm grounded in experimental trial-and-error, heuristic design rules, and incremental optimization. While this approach has yielded transformative technologies, it has also been constrained by the combinatorial vastness of compositional and structural design spaces. Even modest multicomponent systems generate astronomical candidate permutations, rendering exhaustive experimental exploration infeasible. Coupled with the time-intensive nature of synthesis, characterization, and lifecycle validation, traditional discovery pipelines have often required decades to transition from conceptual material to deployable technology.

The emergence of computational materials science has fundamentally reconfigured this paradigm. Early atomistic simulations provided mechanistic insights into structure–property relationships, yet their integration into large-scale discovery workflows remained limited by computational

cost and data fragmentation. A decisive inflection point occurred with the launch of the Materials Genome Initiative in 2011, which articulated a national vision for accelerating materials innovation through the convergence of high-performance computing, digital data infrastructures, and collaborative experimentation [1]. By advocating systematic coupling between computation, data, and experiment, the initiative reframed materials discovery as an informatics-driven enterprise capable of compressing development timelines from decades to years.

Within this emerging paradigm, high-throughput computational screening has become a cornerstone methodology. Automated density functional theory (DFT) workflows now enable the large-scale evaluation of thermodynamic stability, electronic structure, elastic behavior, and defect energetics across thousands to millions of candidate materials [2-4]. These infrastructures transform discovery from artisanal exploration into industrialized simulation, where compositional spaces can be traversed algorithmically rather than experimentally. Such platforms have proven particularly influential in energy materials, superconductors, catalytic systems, and structural alloys, where rapid prioritization of viable candidates significantly reduces experimental burden.

The integration of data-driven modeling has further amplified these capabilities. Machine learning (ML) systems trained on simulation outputs and experimental datasets now provide rapid surrogate predictions for key materials properties, including band gaps, formation energies, diffusion barriers, and mechanical moduli [3, 5-8]. In many contexts, these models approximate quantum-mechanical accuracy while operating at orders-of-magnitude lower computational cost, thereby enabling real-time screening and optimization. Among these architectures, graph neural networks (GNNs) have emerged as particularly powerful due to their natural alignment with atomistic systems. By encoding atoms as nodes and interatomic interactions as edges, GNNs capture local chemical environments and long-range structural dependencies, facilitating transferable predictions across diverse materials classes [5, 9].

Complementing this, representation learning frameworks extract latent descriptors directly from structural data, bypassing handcrafted feature engineering. These embeddings encode hierarchical chemical and topological information, enabling transfer learning between related prediction tasks and supporting cross-domain generalization [5, 10]. Such latent spaces increasingly function as navigable design manifolds, where similarity metrics, clustering, and generative exploration guide discovery trajectories.

High-throughput computational ecosystems typically operate through structured, multi-stage workflows encompassing database construction, descriptor generation, model training, validation, and virtual screening. Open infrastructures such as AFLOW and Materials Project have democratized access to computed materials properties, fostering global collaboration and reproducibility [1, 4]. These repositories aggregate crystallographic, electronic, thermodynamic, and mechanical datasets at unprecedented scale, forming the substrate upon which modern materials informatics operates.

Recent advances have expanded these ecosystems into multimodal data regimes, integrating structural, spectroscopic, imaging, and thermophysical information within unified analytical pipelines [11-13]. Multimodal fusion enhances predictive robustness by embedding complementary physical perspectives into model training. In battery materials research, for instance, informatics frameworks combine electrochemical simulations, microscopy data, and cycling experiments to forecast ion diffusion kinetics, degradation pathways, and stability envelopes [14, 15].

A further inflection point in discovery acceleration is marked by the rise of autonomous materials laboratories. These systems embed machine learning models within closed-loop experimentation pipelines, where active learning algorithms iteratively select high-value experiments based on predictive uncertainty [16-18]. Robotic synthesis, automated characterization, and real-time data assimilation collectively enable self-driving discovery cycles. Thin-film deposition platforms and combinatorial synthesis arrays exemplify this paradigm, demonstrating the capacity to optimize processing parameters and materials compositions with minimal human intervention [19-21].

Parallel to forward screening, inverse design frameworks invert the discovery logic. Rather than predicting properties from known structures, generative models and optimization algorithms propose candidate materials that satisfy predefined functional targets. Variational autoencoders, generative adversarial networks, and reinforcement learning architectures increasingly support this property-to-structure translation, expanding exploration into previously uncharted chemical territories [8, 22].

Despite these transformative advances, systemic challenges persist. Data sparsity in emerging materials classes, model generalization limits, and discontinuities between simulation and experimental realizability continue to constrain predictive reliability. Uncertainty quantification—particularly through Bayesian inference and ensemble modeling—has therefore become integral to adaptive screening workflows, enabling confidence-aware decision-making [4, 16, 17]. Notably, the versatility of these approaches extends beyond crystalline solids; applications in cementitious systems and concrete rheology illustrate the breadth of materials informatics across both hard and soft matter domains [7].

Against this backdrop, the present review examines high-throughput computational screening platforms, data-driven workflows, and their discovery outcomes within contemporary materials engineering. Synthesizing literature, it advances a systems-level interpretation of the field, mapping infrastructural synergies, methodological convergences, and cross-domain translational patterns. By positioning computational screening within broader autonomous and informatics ecosystems, the review bridges foundational simulation paradigms with emergent self-driving discovery architectures, offering strategic insights for scalable materials innovation.

Landscape of Computational & Data-Driven Materials Engineering

Foundations of materials informatics and data ecosystems

Materials informatics constitutes the infrastructural backbone of data-driven discovery, integrating data acquisition, curation, standardization, and analytics into cohesive innovation pipelines [1, 4, 12]. Unlike traditional materials databases—which primarily functioned as passive repositories—modern informatics platforms operate as active knowledge engines, embedding machine learning toolkits, workflow automation, and interoperability protocols.

Large-scale repositories such as the NOMAD Repository exemplify this transformation by aggregating raw simulation outputs and experimental datasets into standardized, AI-ready formats [4]. Through metadata harmonization, ontology development, and descriptor extraction, these ecosystems convert fragmented computational artifacts into actionable discovery intelligence. Standardization frameworks further enable cross-platform interoperability, facilitating collaborative model development and federated data analysis [1, 13].

The evolution toward multimodal materials datasets represents a critical expansion of informatics scope. Integrating crystallographic data with spectroscopy, microscopy, mechanical testing, and thermodynamic profiling allows models to learn richer structure–property correlations [11-13]. For example, coupling scanning transmission electron microscopy with deep learning enables automated defect recognition in 2D materials, linking nanoscale structural irregularities to electronic and mechanical performance [9, 11].

In compositionally complex systems such as high-entropy alloys, informatics strategies leverage statistical redundancy across datasets to compensate for limited labeled data. Machine learning models trained on correlated descriptors can infer phase stability and mechanical resilience despite sparse sampling densities, illustrating how data architecture influences predictive feasibility [6, 10].

Uncertainty quantification is deeply embedded within these ecosystems. Probabilistic modeling frameworks generate confidence intervals around predictions, guiding both data acquisition and experimental prioritization [16, 17]. Active learning infrastructures operationalize this uncertainty by selecting maximally informative samples, thereby optimizing exploration efficiency. Applications include the targeted discovery of high-melting-temperature alloys through coupled machine learning and molecular dynamics simulations, where adaptive sampling reduces computational overhead while preserving predictive accuracy [16, 18]. **Table 1** provides a workflow taxonomy of computational and data-driven materials engineering, mapping platforms, data modalities, model roles, and integration patterns that recur across the reviewed studies.

Table 1. Workflow taxonomy of computational and data-driven materials engineering: platforms, data modalities, ML roles, and integration patterns.

Workflow layer	Primary function	Typical

Data ecosystems	Collect, curate, standardize materials data	Crystal st properties, multimoda images, therr
Representation learning	Learn latent descriptors from raw structures	Atomic gra environment
Property prediction (supervised ML)	Fast surrogate prediction of target properties	Labe simulation/e da
High-throughput screening	Systematic candidate ranking at scale	Large candic
Uncertainty quantification	Quantify confidence; guide exploration	Model output cu
Active learning / acquisition	Choose next computations/experiments	Predictions +
Simulation–experiment integration	Validate + update models with measurements	Rob synthesis/cha lab data
Inverse design	Generate candidates from targets	Property ol const
Translation / deployment	Scale-up and real-world feasibility	Manufacturab const

Machine learning architectures and representation learning

Machine learning now underpins nearly every stage of computational materials engineering, from descriptor extraction to inverse design. Deep learning architectures, in particular, have demonstrated superior performance relative to classical regression and kernel methods, owing to their capacity to model nonlinear, high-dimensional relationships [3, 5, 7, 8, 23].

Graph neural networks remain among the most influential architectures in this domain. Their topology-aware message-passing frameworks encode atomic connectivity, enabling predictive modeling of complex phenomena including defect energetics, catalytic reactivity, and electronic transport [5, 9]. Sparse graph representations further support scalable inference in systems characterized by structural disorder or vacancy distributions.

Representation learning extends these capabilities by constructing hierarchical latent embeddings that capture compositional, structural, and electronic information simultaneously. These learned representations function as universal descriptors, enabling transferability across tasks such as property prediction, clustering, and anomaly detection [8, 10]. In inverse design contexts, generative models operating within these latent spaces can propose structurally plausible yet previously unobserved materials candidates.

Applications span diverse materials classes. In mechanical metamaterials, deep generative frameworks have enabled end-to-end pipelines encompassing structural generation, performance prediction, and optimization [22]. In battery informatics, machine learning architectures integrate electrochemical, structural, and degradation datasets to forecast lifecycle performance and safety envelopes [14, 15].

To address persistent data limitations, researchers increasingly employ redundancy exploitation strategies, wherein correlated descriptors and transfer learning mechanisms enhance predictive performance despite limited labeled datasets [10]. Concurrently, the rise of explainable artificial intelligence (XAI) introduces interpretability layers that elucidate model reasoning

pathways, fostering trust and facilitating physics-consistent validation of predictions [8].

High-throughput computational screening platforms

High-throughput computation involves automated workflows that screen vast libraries of materials using DFT and beyond-DFT methods [2, 3, 19, 24]. Platforms like those for nanoporous materials employ computational screening to identify candidates for targeted applications, such as gas storage, by evaluating properties across thousands of structures [24].

In superconductors, closed-loop approaches combine high-throughput DFT with experimental validation, accelerating discovery through iterative refinement [2]. For organic materials, integrating computational predictions with experimental workflows has streamlined discovery, using ML to prioritize synthesizable candidates [25].

Workflows often incorporate Bayesian optimization to handle experimental failures, as in thin-film growth, where adaptive sampling accounts for variability [17]. DiSCoVeR tool exemplifies screening for unique chemical compositions, focusing on high-performance alloys by filtering databases with ML classifiers [26].

Integration of simulation and experiment

Bridging simulation and experiment is crucial for validating computational predictions [12, 19, 20, 25]. Autonomous laboratories automate this integration, using robotics for synthesis and characterization, guided by ML algorithms [19-21]. In concrete science, ML predicts viscosity and other properties, informing experimental design [7, 27].

Simulation-experiment loops are formalized in active learning frameworks, where models update based on new data to refine predictions [16, 18]. For lithium-ion batteries, ML platforms predict states across development stages, integrating quantum simulations with empirical data [15].

Digital transformation emphasizes data pipelines that connect computational screening to lab automation, as in metamaterials design [28]. The Materials Experiment Knowledge Graph structures this integration, linking entities across studies for holistic analysis [13].

Discovery outcomes and applications

Discovery outcomes from these platforms span multiple domains. In alloys, ML has identified high-melting compositions through active learning and dynamics simulations [18]. For batteries, informatics has uncovered stable electrolytes and cathodes [14, 15].

In 2D materials, sparse ML models predict defect properties, enabling tailored functionalities [9]. Automated experimentation in electron microscopy has revealed atomic-scale insights [11]. Overall, these workflows have led to breakthroughs in energy materials, catalysts, and structural compounds, demonstrating the power of data-driven acceleration.

Autonomous & closed-loop discovery systems

Autonomous discovery systems represent the pinnacle of integration in computational materials engineering, where AI-driven workflows orchestrate the entire discovery process without constant human oversight [12, 19-21]. These systems employ closed-loop architectures that iteratively cycle through hypothesis generation, experimentation, and model refinement, leveraging real-time feedback to optimize outcomes [2, 16, 19, 20].

A core component is active learning, which balances exploration of unknown spaces with exploitation of promising regions [16-18]. In formal terms, this can be conceptualized as an iterative process:

$$\begin{aligned} &Workflow \\ &= argmax x \\ &\in X [\alpha \\ &\cdot u(x) \\ &+ (1 \\ &- \alpha) \\ &\cdot f(x)] \end{aligned} \quad (1)$$

where u(x) represents the uncertainty in prediction for candidate x, f(x) is the predicted fitness, and α \alpha α tunes the exploration-exploitation trade-off. This formulation synthesizes adaptive sampling strategies across studies, guiding selections in high-throughput environments [16, 17].

Closed-loop systems often incorporate robotic platforms for autonomous experimentation, as in materials acceleration platforms that combine AI, high-performance computing, and robotics [19, 21]. For instance, in superconducting

materials, closed-loop workflows have integrated computational screening with synthesis robots, discovering new phases through iterative DFT validations and physical testing [2].

Uncertainty quantification enhances these loops by prioritizing experiments that maximally reduce model variance [4, 16, 17]. Bayesian optimization, resilient to experimental failures, has been applied in materials growth, adjusting parameters in real-time [17]. In electron microscopy, ML automates data acquisition and analysis, closing the loop between imaging and interpretation [11].

Inverse design paradigms within autonomous systems start from target properties and generate structures using generative ML models [8, 22]. For metamaterials, deep learning enables end-to-end inverse workflows, from property specification to fabricable designs [22]. In battery discovery, closed-loops predict and validate ion transport mechanisms, integrating multimodal data [14, 15].

Autonomous laboratories extend this to self-driving setups, where algorithms control synthesis, characterization, and iteration [20, 21]. Community perspectives highlight the need for standardized interfaces to scale these systems [20]. For organic materials, workflows accelerate discovery by linking computational predictions to robotic synthesis [25].

The Materials Experiment Knowledge Graph facilitates knowledge retention across loops, structuring data for reusable insights [13]. In concrete applications, ML estimates viscosity in flow-based processes, enabling autonomous adjustments [7, 27]. **Figure 1** summarizes the end-to-end computational and data-driven materials discovery ecosystem, highlighting how data infrastructures, representation learning, high-throughput screening, and autonomous experimentation couple through uncertainty-aware feedback loops.

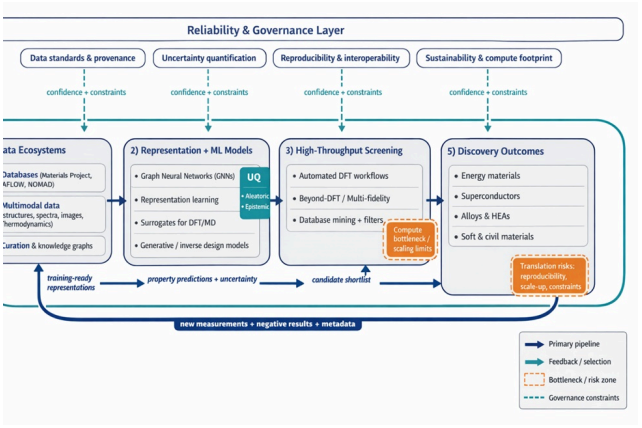

Figure 1. End-to-end ecosystem for computational and data-driven materials engineering.

Schematic overview of the integrated discovery pipeline linking data ecosystems (databases, multimodal inputs, curation), representation learning and machine learning models (including uncertainty quantification), high-throughput computational screening (DFT and multi-fidelity surrogates), and autonomous closed-loop experimentation (active learning–guided synthesis and characterization). Teal feedback channels indicate uncertainty-driven model updates and iterative experimentation; orange envelopes mark scaling bottlenecks and translation risks (compute constraints, interoperability gaps, and reproducibility/scale-up limitations). The spanning governance layer emphasizes standardization, reliability, and sustainability constraints that condition workflow decisions across the pipeline.

Discovery outcomes from these systems include novel alloys via active learning and dynamics [18], and high-performance nanoporous materials through screening [24]. DiSCoVeR tool exemplifies autonomous filtering for unique compositions [26]. AI forecasting extends to predicting research trends, informing loop designs [23].

Overall, autonomous closed-loop systems synthesize computational power with experimental automation, driving efficient materials innovation.

Results and Discussion

The convergence of computational modeling, machine learning, and data-driven infrastructures has reconfigured contemporary materials engineering into an integrated discovery ecosystem. While individual studies report domain-specific advances, a comparative synthesis reveals deeper systemic dynamics shaping scalability,

transferability, and epistemic reliability across workflows [1–29]. Rather than functioning as isolated accelerators, high-throughput simulations, representation learning architectures, and autonomous experimentation platforms now operate as interdependent layers within discovery pipelines, where gains in one layer frequently introduce constraints in another.

A central axis of discussion concerns the scalability of high-throughput screening infrastructures. Cross-domain evidence from superconductors, catalytic systems, and alloy discovery demonstrates that automated density functional theory (DFT) workflows enable rapid traversal of compositional spaces with high electronic-structure fidelity [2, 6, 18, 24]. However, this scalability remains computationally contingent. As chemical and configurational complexity increases—particularly in disordered or multi-principal element systems—computational costs escalate non-linearly, constraining throughput advantages. Moreover, while DFT maintains predictive strength for ground-state energetics and electronic descriptors, it exhibits reduced sensitivity to kinetic pathways, phase transitions, and temperature-dependent dynamics. Hybridization with molecular dynamics (MD) and thermodynamic sampling frameworks therefore emerges as a necessary integrative strategy to capture temporal and stochastic behaviors absent in static approximations [2, 18].

Representation learning constitutes another unifying thread across the reviewed studies. Graph neural networks (GNNs) and related geometric deep learning architectures demonstrate strong capacity for encoding structure–property relationships into transferable latent embeddings spanning crystalline solids, 2D materials, and architected metamaterials [5, 9, 22]. These embeddings enable cross-property inference and accelerated screening; however, comparative analysis reveals that representation robustness is highly contingent on dataset topology. Heterogeneous datasets—particularly those integrating experimental spectra, simulation outputs, and imaging modalities—introduce fusion asymmetries that complicate model optimization. For example, multimodal battery informatics frameworks improve predictive granularity for degradation and performance metrics, yet they require complex alignment protocols to reconcile temporal, spatial, and physicochemical scales embedded within the data streams [12-15].

Active learning systems further operationalize representation infrastructures by steering data acquisition through uncertainty-aware sampling. Synthesized evidence suggests that Bayesian optimization and ensemble-driven acquisition functions significantly reduce experimental search burdens [16, 17]. Nevertheless, their performance is domain-sensitive. In thin-film synthesis and growth optimization, experimental noise, stochastic defects, and process instabilities attenuate the reliability of uncertainty estimates, leading to suboptimal sampling trajectories and increased iteration cycles. Thus, while active learning enhances exploration efficiency, its epistemic strength remains tethered to experimental signal quality.

Autonomous laboratories integrate high-throughput computation, machine learning inference, and robotic experimentation into closed-loop discovery architectures. These self-driving systems demonstrate the potential to compress design–synthesis–characterization cycles substantially [19-21]. Yet interoperability remains a critical friction point. Cross-study comparisons highlight mismatches between simulation parameterizations and experimental execution constraints, compounded by the absence of standardized communication protocols and application programming interfaces (APIs). Consequently, translation gaps persist between predicted optima and experimentally reproducible materials outcomes.

Domain-specific applications—including concrete informatics, polymer design, and organic electronics—illustrate both the breadth and the limits of data-driven generalization. While predictive models accelerate formulation optimization and property forecasting, they often struggle with high-dimensional compositional manifolds characterized by sparse sampling densities [7, 9, 25, 27]. Dimensionality reduction and sparse encoding techniques partially mitigate overfitting risks but do not eliminate representation blind spots, particularly in extrapolative regimes.

Collectively, the synthesized literature frames computational materials engineering not merely as a toolkit of acceleration technologies but as a workflow-centric epistemic system. Discovery outcomes are increasingly shaped by how data infrastructures, models, and automation platforms are architected and integrated. Achieving durable progress therefore requires calibrated balancing between computational efficiency, representational fidelity, and empirical validation.

Challenges & limitations

Despite rapid methodological maturation, several structural bottlenecks continue to constrain the full realization of computationally driven materials discovery ecosystems [1, 3, 5, 8, 12, 20].

Data scarcity remains a foundational limitation. While flagship databases have expanded coverage for crystalline compounds, niche domains—such as high-entropy alloys, metastable phases, and extreme-environment materials—remain sparsely characterized. Redundant exploitation strategies, transfer learning, and generative augmentation partially alleviate these deficits, yet they cannot fully substitute for experimentally validated ground truth, particularly where compositional disorder or kinetic stabilization dominates [6, 10].

Computational cost escalation further limits scalability. High-throughput screening campaigns targeting nanoporous frameworks, catalytic surfaces, or multicomponent alloys often require evaluation of millions of candidate structures, imposing prohibitive demands on compute infrastructure and energy consumption [24]. As model complexity and configurational resolution increase, throughput gains risk being offset by resource intensiveness.

Model generalizability presents another critical constraint. Machine learning systems trained predominantly on ordered crystalline datasets frequently underperform when extended to amorphous, defective, or hybrid organic–inorganic systems. Defect energetics, grain boundaries, and disorder introduce representational discontinuities that challenge learned embeddings and reduce predictive reliability [5, 7, 9].

Closely related are limitations in uncertainty quantification. Although Bayesian neural networks, ensemble modeling, and probabilistic inference frameworks have advanced confidence estimation, they remain less robust in extrapolative chemical spaces where epistemic uncertainty dominates [4, 16, 17]. Miscalibrated uncertainty can propagate risk across downstream screening and optimization workflows.

Autonomous laboratory infrastructures face both technical and operational constraints. Hardware throughput ceilings, robotic dexterity limits, and integration barriers between wet-lab instrumentation and digital control systems restrict experimental scalability [19-21]. Furthermore, discrepancies between simulated synthesis pathways and real-world process constraints introduce executional drift in closed-loop systems.

Multiscale integration remains an unresolved grand challenge. Bridging quantum-scale simulations with mesoscopic microstructures and macroscopic performance metrics requires harmonized multimodal datasets and advanced fusion algorithms [11-13, 15]. In battery informatics, for instance, lifecycle and degradation predictions are hindered by insufficient longitudinal datasets capturing dynamic electrochemical evolution [14, 15].

Experimental practice introduces additional friction through noise, failure rates, and reproducibility variance. Materials growth processes—thin films, crystals, and additive manufacturing—exhibit stochastic sensitivities that complicate optimization and inflate iteration costs [17].

Finally, ethical, infrastructural, and sustainability considerations are gaining visibility. Data-sharing asymmetries across institutional and industrial ecosystems impede collaborative model development, while the environmental footprint of large-scale computation raises questions regarding the sustainability of ever-expanding screening campaigns [1, 4, 28].

Synthesis outlook

Addressing these challenges will require coordinated advances across data generation, algorithm design, and experimental integration. Hybrid physics–AI models, federated data ecosystems, energy-aware computing strategies, and standardized automation interfaces represent critical pathways forward. Only through such systemic alignment can computational materials engineering transition from accelerated discovery toward truly resilient and generalizable innovation infrastructures.

Future research directions

Future trajectories in computational and data-driven materials engineering should prioritize scalable, integrative systems to overcome current limitations [1, 8, 12, 23, 28, 29]. One direction involves advancing multimodal data fusion with AI toolkits, enabling seamless simulation-experiment loops for emergent properties in complex materials [4, 12, 13]. Enhancing active learning with multi-fidelity models could optimize resource use, incorporating

low-cost surrogates for initial screening before high-fidelity DFT [16, 18].

In autonomous discovery, developing modular platforms with plug-and-play robotics will facilitate broader adoption, as advocated in community roadmaps [19-21]. Inverse design should evolve towards multi-objective optimization, formalized as:

$$\begin{array}{l} Design \\ = argminx \sum iwi \\ \cdot Li\left(f\left(x\right), yi\right) \\ + \lambda \\ \cdot c\left(x\right) \end{array} \quad (2)$$

where L_i are loss terms for properties y_i, weights w_i, and c(x) constraints like synthesizability, synthesizing cross-study needs for balanced workflows [8, 22].

Uncertainty-aware generative models promise breakthroughs in unexplored chemical spaces, particularly for sustainable materials [17, 29]. In battery and concrete domains, real-time adaptive systems could predict long-term behaviors by integrating time-series data [7, 14, 15, 27].

Digital twins for materials ecosystems, linking knowledge graphs with predictive analytics, will enable predictive maintenance and trend forecasting [13, 23, 28]. Collaborative infrastructures, emphasizing open-source tools, are crucial for democratizing access [1, 4]. These directions position the field for transformative impacts in energy, electronics, and beyond.

Conclusion

Computational and data-driven materials engineering has matured into a powerful paradigm, with high-throughput platforms and closed-loop workflows driving unprecedented discovery rates. By synthesizing informatics, ML, and autonomous systems, the field addresses complex challenges through integrated ecosystems. Future advancements will hinge on resolving data and integration hurdles, paving the way for inverse-designed materials with tailored functionalities. This review underscores the transformative potential, advocating for continued innovation in workflow synergies to accelerate materials breakthroughs.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 31 Dec 2022 Revised: 28 Mar 2023 Accepted: 31 May 2023
Published online: 18 September 2023

References

de Pablo JJ, Jackson NE, Webb MA, Chen L-Q, Moore JE, Morgan D, et al. New frontiers for the materials genome initiative. npj Comput Mater. 2019;5(1):41. https://doi.org/10.1038/s41524-019-0173-4.

Dan Y, Zhao T, Adu-Yeboah F, Brockschnieder S, Lydic R, McQuade D, et al. Closed-loop superconducting materials

discovery. npj Comput Mater. 2023;9(1):131.

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(1):83.

Ghiringhelli LM, Carbogno C, Levchenko S, Mohamed F, Huhs G, Scheffler M, et al. The NOMAD artificial-intelligence toolkit: Turning materials-science data into knowledge and understanding. npj Comput Mater. 2022;8(1):258. https://doi.org/10.1038/s41524-022-00935-z.

Choudhary K, DeCost B, Chen C, Jain A, Tavazza F, Cohn R, et al. Recent advances and applications of deep learning methods in materials science. npj Comput Mater. 2022;8(1):59. https://doi.org/10.1038/s41524-022-00734-6.

Zhou Z, Zhou Y, He Q, Ding Z, Li FC, Yang Y. Machine learning guided appraisal and exploration of phase design for high entropy alloys. npj Comput Mater. 2019;5(1):64.

Geng Z, Geitner M, Tan F, Himmelreich J, Lorke A, Tremblay JC. Machine learning in concrete science: Applications, challenges, and best practices. npj Comput Mater. 2022;8(1):127. https://doi.org/10.1038/s41524-022-00810-x.

Pilania G, Gubernatis JE, Lookman T. Machine learning in materials science: From explainable predictions to autonomous design. Comput Mater Sci. 2021;193:110360

Bihlmayer G, Schmidt SJ, Henkelman G, Franchini C. Sparse representation for machine learning the properties of defects in 2D materials. npj Comput Mater. 2023;9(1):31. https://doi.org/10.1038/s41524-023-01062-z.

Chen C, Zuo Y, Ye W, Li X, Ong SP. Exploiting redundancy in large materials datasets for efficient machine learning with less data. Nat Commun. 2023;14(1):7286. https://doi.org/10.1038/s41467-023-42992-y.

Ghosh A, Ziatdinov M, Nelson CT, Vasudevan RK, Kalinin SV. Machine learning for automated experimentation in scanning transmission electron microscopy. npj Comput Mater. 2023;9(1):184. https://doi.org/10.1038/s41524-023-01142-0.

Liu Y, Guo B, Zou Z, Xiao Y, Lei X, Song D, et al. AI applications through the whole life cycle of material discovery. Matter. 2020;3(4):980-1000.

De Almeida AF, Moreira E, Rodrigues RP, Li J, Abranches DO, Costa JCS, et al. The materials experiment knowledge graph. Digit Discov. 2023;2(3):633-47.

Wang L, Chen S, Li W, Ren F, Wang Y, Tan Z, et al. A review of the recent progress in battery informatics. npj Comput Mater. 2022;8(1):66. https://doi.org/10.1038/s41524-022-00713-x.

Lv C, Zhou X, Zhong L, Min C, Shi M, Ma X, et al. Machine learning: An advanced platform for materials development and state prediction in lithium-ion batteries. Adv Mater. 2022;34(42):2101474.

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(1):21. https://doi.org/10.1038/s41524-019-0153-8.

Shimizu R, Kobayashi S, Watanabe Y, Ando Y, Hitosugi T. Bayesian optimization with experimental failure for high-throughput materials growth. npj Comput Mater. 2022;8(1):175. https://doi.org/10.1038/s41524-022-00859-8.

Divilov S, Eckert H, Majzoub EH, Wolverton C. Active learning and molecular dynamics simulations to find high melting temperature alloys. Comput Mater Sci. 2022;208:111292.

Abolhasani M, Kumara Alagiyawanna KAS. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. npj Comput Mater. 2022;8(1):89. https://doi.org/10.1038/s41524-022-00765-z.

Stach E, DeCost B, Kusne AG, Hattrick-Simpers J, Brown KA, Reyes KG, ET AL. Autonomous experimentation systems for materials development: A community perspective. Matter. 2021;4(9):2702-26.

Christensen M, Yunker LPE, Shiran H, Adedeji F, Gormley AJ, Drobot B, et al. Integrating autonomy into automated research platforms. Digit Discov. 2023;2(6):1644-53.

Zheng J, Phipps RJ. Deep learning in mechanical metamaterials: from prediction and generation to inverse design. Adv Mater. 2023;35(49):2302530.

Krenn M, Pollice R, Guo SY, Aldeghi M, Cervera-Lierta A, Friederich P, et al. Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network. Nat Mach Intell. 2023;5(11):1326-35. https://doi.org/10.1038/s42256-023-00735-0.

Erlekam F, Feng X, Lan G, Johnson T, Gopakumar A, Farha OK, et al. High-throughput computational screening of

nanoporous materials in targeted applications. Digit Discov. 2022;1(6):706-23.

Greenaway RL, Kontoravdi C, Kim B, Livingstone K, Campbell AJ, Adjiman CSJ. Integrating computational and experimental workflows for accelerated organic materials discovery. Adv Mater. 2021;33(11):2004831.

Baird SG, Tran TD, Liu Z, Fan W, Sparks TD. DiSCoVeR: A materials discovery screening tool for high performance, unique chemical compositions. Digit Discov. 2022;1(3):226-40.

Ahlhelm M, Günther A, Carneiro OS, Milewski A, Günther S, Riedel R, et al. Go with the flow: Deep learning methods for autonomous viscosity estimations. Digit Discov. 2023;2(6):1672-85.

Scheideler WJ, McRae O. Digital transformation in materials science: A paradigm change in material's development. Adv Mater. 2021;33(15):2007940.

López Lorente ÁI, González-Fernández S, Cerdán-Pasarán A. Artificial Intelligence and Advanced Materials. Adv Mater. 2023;35(40):2208683.