ORIGINAL RESEARCH **Open access**

Scaling Discovery Infrastructures: Computational Resource Allocation in Materials Engineering and Computational Data

Carlos Vega[1*], Maria Hernandez[1]

Abstract

In the evolving landscape of computational and data-driven materials engineering, the integration of advanced machine learning techniques with high-throughput simulations has transformed discovery pipelines, enabling accelerated identification of novel materials. However, as datasets grow in multimodality and scale, and models incorporate complex architectures such as graph neural networks, the allocation of computational resources emerges as a critical bottleneck. This conceptual manuscript addresses the infrastructural challenges in scaling these ecosystems, highlighting gaps in resource orchestration that hinder efficient coupling of simulation, experimentation, and inference processes. We introduce a novel framework, termed the Adaptive Resource Equilibrium Model (AREM), which conceptualizes resource allocation as a dynamic interplay between data representation fidelity, model computational demands, and discovery throughput. By synthesizing insights from materials informatics and autonomous systems, AREM emphasizes feedback mechanisms to balance epistemic uncertainties and infrastructural constraints, fostering resilient discovery infrastructures. The implications extend to enhancing inverse design workflows and closed-loop experimentation, potentially streamlining resource utilization in large-scale materials research consortia. This work provides a systems-level perspective on optimizing computational ecosystems, guiding future developments in scalable, data-centric materials engineering without empirical validation.

Keywords Materials informatics, Uncertainty quantification, Machine learning architectures, Computational resource allocation, High-throughput discovery, Autonomous feedback loops

*Correspondence:
Carlos Vega
carlos.vega@gmail.com
[1] Department of Computational Materials Science, Faculty of Engineering, Polytechnic University of Valencia, Valencia, Spain

Introduction

The Paradigm Shift in Materials Discovery The field of materials engineering has undergone a profound transformation with the advent of computational and data-driven methodologies. Traditionally reliant on empirical trial-and-error approaches, materials research now leverages vast datasets and sophisticated algorithms to predict properties, design structures, and accelerate innovation. This shift is underpinned by advancements in materials informatics, where data from simulations and experiments are harnessed to inform predictive models [1, 2]. High-throughput computation, for instance, enables the screening of thousands of candidate materials in silico, reducing the time and cost associated with physical synthesis [3, 4]. Concurrently, machine learning techniques, including representation learning and deep architectures, have become integral to interpreting complex material behaviors, from bandgap predictions to polymer property forecasts [5-8].

Yet, as these methods proliferate, the scale of computational demands escalates. Multimodal datasets—encompassing structural, electronic, and thermodynamic

information—require substantial processing power, often distributed across heterogeneous infrastructures [9, 10]. Foundation models for science, drawing from large-scale pretraining, further amplify this need by integrating diverse data modalities [11]. In this context, resource allocation emerges not merely as a logistical concern but as a foundational element influencing the efficacy of discovery pipelines. Inefficient allocation can lead to bottlenecks in simulation-experiment coupling, where delays in one stage propagate through the entire workflow [12, 13].

Challenges in Scaling Computational Ecosystems Scaling discovery infrastructures involves more than increasing hardware capacity; it necessitates intelligent orchestration of resources to align with the dynamic nature of data-driven tasks. For example, graph neural networks, widely applied in crystal and molecular modeling, demand specialized computational environments that balance memory usage with parallel processing [10, 14, 15]. Uncertainty quantification adds another layer, as models must account for epistemic and aleatory variabilities, often requiring iterative refinements that consume additional cycles [16, 17]. Autonomous discovery systems, which incorporate closed-loop experimentation, exacerbate these issues by introducing real-time feedback that requires adaptive resource provisioning [4, 13].

Literature highlights persistent gaps in this area. While high-throughput frameworks have democratized access to computational tools [18, 19], they often overlook the interplay between data volume and resource efficiency [20, 21]. Inverse materials design, aiming to engineer materials from desired properties backward, relies on generative models like GANs, yet their training phases can overwhelm standard infrastructures without tailored allocation strategies [20, 22]. Moreover, as research ecosystems expand to include collaborative networks, disparities in resource availability across institutions pose equity challenges, potentially skewing discovery outcomes [23, 24].

Toward Infrastructural Resilience Addressing these challenges requires a conceptual reevaluation of how resources are allocated within computational materials engineering. Rather than viewing resources as static assets, they should be treated as adaptive components responsive to workflow demands. This perspective draws from systems engineering principles, emphasizing modularity and scalability [25, 26]. Data-driven paradigms, enriched by multimodal datasets, offer opportunities for smarter allocation through predictive analytics on resource needs [9, 27].

However, existing approaches fall short in providing a unified framework that integrates these elements. Many focus on isolated aspects, such as model optimization or dataset curation, without considering holistic infrastructure dynamics [28, 29]. This manuscript positions a new conceptual framework to bridge this gap, conceptualizing resource allocation as an equilibrium process that harmonizes data fidelity, computational intensity, and discovery velocity. By doing so, it aims to enhance the resilience and efficiency of materials discovery infrastructures.

Theoretical Background & Literature Synthesis

Evolution of computational paradigms in materials engineering

From first-principles physics to computational screening

The foundational transformation toward computational materials science emerged from the maturation of quantum-mechanical simulation frameworks, particularly density functional theory (DFT), which enabled the prediction of structural, electronic, and thermodynamic properties directly from atomic configurations [5, 18]. These first-principles approaches established the feasibility of *in silico* materials characterization, reducing reliance on trial-and-error experimentation and enabling predictive insights prior to synthesis.

Early computational paradigms were primarily physics-driven, emphasizing mechanistic fidelity over scale. Simulations were computationally expensive, limiting exploration to narrow chemical subspaces. However, the integration of machine learning (ML) into atomistic modeling workflows began to alleviate these constraints by introducing surrogate models capable of approximating DFT outputs at significantly reduced computational cost [5, 18].

Emergence of materials informatics ecosystems

Over the past decade, computational materials science has evolved into a data-intensive informatics discipline. Large-scale repositories—spanning crystallographic databases,

thermodynamic archives, and spectroscopy datasets—have enabled the aggregation of multimodal knowledge infrastructures [2, 19]. This transition marks a shift from isolated simulations to interconnected data ecosystems in which models are trained on aggregated experimental and computational corpora.

High-throughput computation has been central to this transformation. Automated workflows now enable systematic traversal of compositional and structural design spaces across alloys, ceramics, polymers, and energy materials [3, 4, 21]. Parallelized simulation architectures facilitate exponential scaling, allowing millions of candidate materials to be screened through distributed computing pipelines [15, 25].

Representation learning and structural encoding

Machine learning has further reshaped computational paradigms through advances in representation learning. Materials are no longer described solely through handcrafted descriptors; instead, they are encoded as latent vectors, tensors, or graph structures that preserve atomic topology and bonding environments [6, 10].

Graph neural networks (GNNs) exemplify this paradigm shift. By modeling atoms as nodes and interatomic interactions as edges, GNNs capture relational dependencies critical to predicting formation energies, phase stability, and electronic properties [12, 14]. Such architectures enable end-to-end learning from raw structural data, reducing reliance on domain-engineered features.

From rule-based simulation to data-centric discovery

Collectively, these developments signal a broader epistemic transition—from rule-based simulation toward data-centric inference systems. Discovery increasingly emerges from statistical pattern recognition across vast datasets rather than solely from mechanistic modeling [1, 23]. This evolution expands discovery throughput but introduces new dependencies on data quality, representational adequacy, and computational scale.

Data-driven infrastructures and resource demands

Multimodal data ecosystems

Contemporary materials engineering infrastructures are underpinned by multimodal datasets integrating simulation outputs with experimental validation streams [9, 11]. These datasets span spectroscopic signatures, mechanical response profiles, microstructural imaging, and thermodynamic measurements, forming high-dimensional knowledge matrices that capture materials behavior across scales.

The integration of heterogeneous modalities enhances predictive robustness but imposes significant demands on data harmonization, preprocessing, and feature fusion pipelines [27]. Storage architectures must accommodate petascale data volumes, while processing frameworks require high-performance computing (HPC) orchestration.

Foundation models for materials science

Recent advances in scientific foundation models introduce transfer learning paradigms capable of generalizing knowledge across material classes [11]. Pretrained on vast simulation and experimental corpora, these models enable downstream fine-tuning for property prediction, inverse design, and anomaly detection tasks.

However, the training of such models is computationally intensive. Distributed GPU clusters, large-memory nodes, and high-bandwidth data pipelines are often required, amplifying infrastructural asymmetries between well-resourced and resource-constrained research environments [8, 24].

Computational resource stratification

Resource demands are further compounded by uncertainty quantification (UQ) requirements. Ensuring model reliability in safety-critical domains—such as battery electrolytes or catalytic materials—necessitates probabilistic inference frameworks [16, 17]. Ensemble modeling, Bayesian neural networks, and Monte Carlo dropout approaches propagate predictive uncertainty but multiply computational overhead.

Bayesian active learning provides one mitigation strategy by steering simulations toward high-information regions of compositional space [13, 17]. Acquisition functions prioritize candidates with maximal epistemic uncertainty, optimizing computational expenditure.

Autonomous discovery and resource allocation

Autonomous discovery platforms extend these principles by embedding AI agents within simulation–experiment loops

[4, 13]. These systems dynamically allocate computational and experimental resources, iteratively refining hypotheses with minimal human intervention.

Yet, autonomy does not eliminate resource constraints. Fixed computational budgets, laboratory throughput limits, and instrumentation bottlenecks often result in uneven resource distribution across pipeline stages [7, 20]. Consequently, efficiency gains at one layer may induce bottlenecks elsewhere, revealing systemic allocation imbalances.

Integration of simulation and experimentation

Closed-loop discovery architectures

The coupling of computational simulations with experimental validation constitutes a cornerstone of contemporary materials discovery infrastructures [12, 13]. Closed-loop frameworks integrate predictive modeling, synthesis, characterization, and feedback analysis into iterative optimization cycles.

High-throughput experimentation accelerates this process by enabling rapid synthesis and screening of candidate materials [9]. Machine learning diagnostic tools interpret experimental outputs in real time, informing subsequent simulation cycles.

Inverse design and generative modeling

Inverse design paradigms invert traditional forward modeling. Rather than predicting properties from known materials, generative systems propose candidate structures that satisfy predefined functional targets [20, 22].

Deep generative architectures—including generative adversarial networks (GANs) and variational autoencoders (VAEs)—sample latent chemical spaces to identify viable compositions [7, 20]. These models enable efficient navigation of high-dimensional design landscapes, accelerating the identification of high-performance materials.

Translational bottlenecks across platforms

Despite integrative advances, infrastructural discontinuities persist. Simulation environments, robotic synthesis platforms, and characterization systems often operate on heterogeneous software and hardware stacks [18, 26]. Data interoperability challenges and workflow incompatibilities introduce latency between computational prediction and physical validation.

Empirical domains such as perovskite photovoltaics and high-entropy alloys illustrate these translational frictions, where uneven investment across simulation and experimental layers constrains discovery throughput [4, 9, 29].

Multimodal representation fusion

As discovery pipelines integrate imaging, spectroscopy, and atomistic simulations, representation learning must accommodate multimodal fusion [10, 11]. Cross-modal embeddings and co-attention architectures enable joint reasoning across heterogeneous data streams but substantially increase computational and architectural complexity.

Epistemic and systemic considerations

Bias, incompleteness, and epistemic risk

At the systems level, discovery infrastructures must contend with epistemic vulnerabilities. Training datasets frequently reflect historical experimental biases, compositional feasibility constraints, and reporting asymmetries [16, 23]. Models trained on such datasets risk extrapolation errors when deployed in under-sampled chemical domains.

Uncertainty quantification frameworks mitigate these risks by propagating predictive variance across workflows. However, ensemble simulations and probabilistic inference introduce additional computational burdens [14, 17].

Collaborative infrastructures and data commons

Open scientific databases—such as large quantum materials repositories—facilitate collaborative discovery by democratizing access to computational datasets [19, 27]. These platforms accelerate benchmarking, reproducibility, and cross-institutional research.

Yet, access to computational resources required to leverage these datasets remains uneven. Infrastructure disparities create asymmetries in model development capacity, influencing global discovery participation.

Orchestration gaps in discovery ecosystems

Literature synthesis reveals a critical systems gap: while individual components—datasets, models, and algorithms

—have matured significantly, their orchestration within scalable infrastructures remains underdeveloped [2, 21, 25]. Integration inefficiencies manifest as data silos, workflow discontinuities, and computational redundancies.

Accuracy–efficiency trade-offs

Discovery infrastructures must continuously negotiate trade-offs between predictive fidelity and computational efficiency. Complex architectures yield deeper insights but demand greater training time, memory allocation, and energy consumption [5, 6, 8].

Autonomous systems and steering logics

Autonomous laboratories offer partial mitigation through adaptive steering mechanisms that dynamically allocate resources across discovery stages [4, 12, 13]. However, absent cohesive infrastructural governance frameworks, such systems risk amplifying rather than resolving inefficiencies.

Synthesis: Toward resource-aware discovery infrastructures

Cross-domain synthesis reveals that computational materials engineering has entered an infrastructural phase characterized by scale, heterogeneity, and autonomy. Discovery outcomes are no longer determined solely by algorithmic performance but by the orchestration of computational, experimental, and epistemic resources.

Persistent tensions emerge:

- Scalability vs interpretability
- Predictive accuracy vs computational cost
- Data abundance vs representational bias
- Autonomy vs governance

These tensions foreground resource allocation as a central systems variable rather than a peripheral operational concern. Conceptualizing resource distribution as dynamic, adaptive, and epistemically informed provides a foundation for next-generation discovery infrastructures capable of balancing computational demand with scientific insight [1, 10]. The stratification of computational paradigms and their corresponding infrastructural demands are synthesized in **Table 1**.

Table 1. Computational Paradigm Layers and Resource Demands in Data-Driven Materials Discovery

Paradigm Layer	Core Functions	Dominant Data Types	Computation Demands
First-Principles Simulation	Atomistic property prediction	Quantum mechanical outputs	Extreme HP compute
High-Throughput Screening	Composition space exploration	Structured simulation datasets	Parallel compute clusters
Materials Informatics	Data aggregation & modeling	Multimodal repositories	Distributed processing
Representation Learning	Structural encoding	Graphs, embeddings	GPU trainin loads
Foundation Models	Transferable inference	Large multimodal corpora	Pretraining megascale compute
Autonomous Discovery	Closed-loop optimization	Real-time multimodal streams	Continuous compute allocation

Proposed conceptual framework

The Adaptive Resource Equilibrium Model (AREM) To address the infrastructural challenges in scaling discovery ecosystems, we propose the Adaptive Resource Equilibrium Model (AREM), an original framework that conceptualizes computational resource allocation as a dynamic equilibrium process within materials engineering pipelines. AREM structures the discovery infrastructure into three interconnected layers: the Data Representation Layer, the Model Inference Layer, and the Discovery Steering Layer. These layers interact through feedback loops that adjust resource distribution based on real-time workflow dynamics, ensuring alignment between data complexity, computational intensity, and output throughput.

At the core, the Data Representation Layer handles multimodal inputs, transforming raw datasets into encoded forms suitable for downstream processing. This layer emphasizes fidelity in representations, such as graph-based encodings for crystalline structures or vector embeddings for polymeric chains, while monitoring

resource consumption during feature extraction. Resources here are allocated proportionally to data heterogeneity, with feedback from higher layers signaling adjustments if epistemic uncertainties arise from incomplete modalities.

Transitioning upward, the Model Inference Layer deploys architectures like graph neural networks or generative models to perform predictions and optimizations. This layer captures the bulk of computational demands, where inverse design tasks or uncertainty quantifications require iterative computations. AREM introduces equilibrium logic to balance model depth with resource availability, preventing overloads in high-throughput scenarios.

The Discovery Steering Layer oversees the overall pipeline, integrating simulation-experiment couplings and autonomous loops. It employs steering logics to redistribute resources, such as prioritizing closed-loop iterations when discovery velocity dips below equilibrium thresholds. Feedback loops connect all layers: upward flows propagate uncertainty signals, while downward adjustments optimize allocations, fostering resilience against scaling bottlenecks. The layered architecture and feedback-driven allocation logic of AREM are illustrated in **Figure 1**, which conceptualizes computational discovery infrastructures as equilibrium systems balancing data complexity, inference intensity, and discovery throughput through adaptive resource steering.

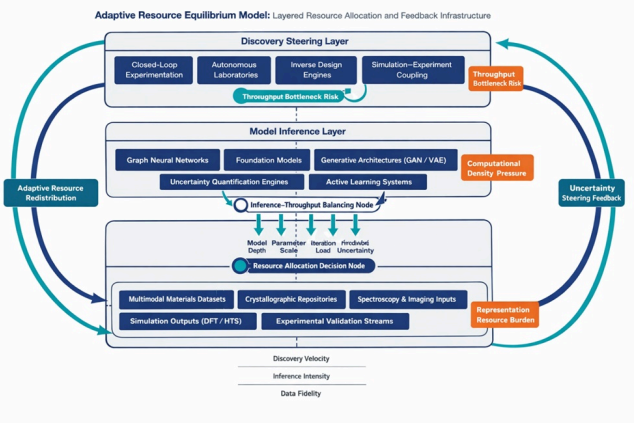

Figure 1. Adaptive Resource Equilibrium Model (AREM) for scalable computational materials discovery infrastructures.

The framework conceptualizes resource allocation as a dynamic equilibrium process across three interconnected layers: Data Representation, Model Inference, and Discovery Steering. Equilibrium nodes regulate resource distribution based on data complexity, model intensity, and throughput demands. Bidirectional feedback loops propagate uncertainty and performance signals, enabling adaptive redistribution of computational capacity across simulation, learning, and experimental workflows. The model illustrates how infrastructural resilience emerges from balancing epistemic fidelity with computational scalability.

To formalize the interaction between data fidelity and resource demands, the equilibrium in AREM may be expressed as $R_{eq} = \frac{f(D_c * M_i)}{S_v}$, where R denotes the equilibrium resource state, D_c symbolizes data complexity (e.g., modality count and volume), M_i represents model inference intensity (e.g., parameter scale and iteration depth), and S_v captures steering velocity (e.g., pipeline throughput rate). This expression captures the interaction between input demands and output efficiency, guiding allocations to maintain systemic balance without empirical metrics.

Further, the feedback dynamics in autonomous loops can be conceptualized as $\Delta R = \alpha \ (U_e - U_t)$, where ΔR is the resource adjustment increment, U_e is the estimated epistemic uncertainty from the inference layer, U_t is a threshold uncertainty level, and α is a scaling factor reflecting infrastructure elasticity. This formula highlights how deviations in uncertainty trigger reallocations, enhancing adaptability in closed-loop systems.

Finally, the trade-off between representation accuracy and discovery speed may be expressed as $T_o = \beta \left(\frac{R_f}{D_f}\right) + \gamma \left(\frac{1}{C_d}\right)$, where T_o is the overall trade-off operator, R_f is representation fidelity, D_f is data fusion efficiency, C_d is computational density, and β, γ are weighting coefficients for layer-specific contributions. This captures the systemic tensions, interpreting how optimizations in one layer influence others.

Through these elements, AREM provides a conceptual blueprint for scalable infrastructures, integrating computational workflows with epistemic considerations in materials engineering.

Analytical implications

The Adaptive Resource Equilibrium Model (AREM) offers interpretive insights into the dynamics of computational materials engineering, particularly in how resource allocation influences pipeline efficiency and epistemic robustness. By framing infrastructures as layered systems with feedback mechanisms, AREM elucidates trade-offs inherent in scaling discovery processes. For instance, in high-throughput computation, where vast composition spaces are explored [3, 4, 21], the model's equilibrium logic interprets how over-allocation to the Data Representation Layer can diminish overall throughput if not balanced against Model Inference demands. This perspective reveals that data complexity, often amplified by multimodal integrations [9, 11, 27], interacts with resource constraints to shape inference accuracy, without implying empirical thresholds.

Consider the implications for machine learning architectures in materials informatics. Graph neural networks, adept at capturing structural dependencies [10, 14, 15], impose variable computational loads based on graph size and depth. AREM interprets this as a tension between representation fidelity and inference speed, where feedback loops can steer resources toward uncertainty hotspots [16, 17]. In inverse design scenarios, generative models like those based on adversarial networks [7, 20] benefit from this steering, as AREM conceptualizes resource redistribution to favor exploratory sampling over redundant validations [22]. Such dynamics highlight infrastructural resilience, interpreting how adaptive allocations mitigate risks in autonomous systems, where closed-loop experimentation demands real-time responsiveness [4, 12, 13].

Epistemic risk structures are another key area. Uncertainty quantification, integral to reliable predictions [5, 6, 16], is interpreted through AREM as a feedback-driven process that allocates resources proportionally to variance estimates. This avoids static provisioning, instead fostering interactions where simulation-experiment couplings [9, 12, 18] are optimized for epistemic closure. For foundation models in science [11], the framework implies that pretraining phases, resource-intensive due to large datasets [8, 24], can be equilibrated by downscaling during fine-tuning, preserving discovery velocity.

In collaborative ecosystems, AREM provides systems-level insights into resource equity. Open databases and shared workflows [19, 27] often exhibit disparities, and the model interprets these as disequilibria that feedback loops can address through predictive steering [25, 26, 28]. This extends to polymer design and energy materials, where machine-driven property predictions [6-8] intersect with infrastructural limits, implying that balanced allocations enhance generalizability across material classes [1, 2, 23].

To capture the interplay of these factors, the resource-dynamics trade-off in AREM can be conceptualized as $E_d = \delta \left(\frac{R_a}{U_p}\right) - \varepsilon \ (C_v)$, where E_d denotes epistemic discovery efficiency, R_a is allocated resources, U_p is propagated uncertainty, C_v is computational variance across layers, and δ, ε are interpretive coefficients for amplification and damping effects. This expression interprets how uncertainty mitigation amplifies efficiency while variance dampens it, guiding conceptual optimizations in discovery steering.

Furthermore, the scalability implication may be expressed as $S_i = \zeta \int \frac{(F_l dt)}{R_c}$ where S_i is infrastructure scalability index, F_l represents feedback loop frequency over time t, R_c is resource capacity, and ζ is a normalization factor. This captures the cumulative effect of feedbacks on scaling, interpreting sustained equilibrium as a driver for expansive ecosystems.

The equilibrium interactions governing resource redistribution across AREM layers are systematized in **Table 2**.

Table 2. AREM Resource Equilibrium Dynamics Across Discovery Infrastructure Layers

AREM Layer	Resource Inputs	Allocation Drivers	Feedback Signals
Data Representation	Multimodal datasets, repositories	Modality volume, encoding cost	Data sparsity, fusion error
Model Inference	Compute cycles, GPU clusters	Model depth, parameter scale	Predictive uncertainty
Discovery Steering	Experimental capacity, orchestration systems	Throughput targets, design priorities	Discovery velocity gaps

Cross-Layer Interfaces	Shared computational pools	Task concurrency	Variance propagation
Autonomous Feedback Loops	Real-time telemetry	Uncertainty thresholds	Epistemic deviation

These analytical implications underscore AREM's role in reinterpreting computational workflows, offering a lens for balancing demands in data-driven materials engineering [10, 29].

Results and Discussion

The conceptualization of AREM within computational materials engineering invites a broader examination of its integrative potential and limitations in current paradigms. While the framework emphasizes dynamic equilibrium, it interprets the challenges of implementing such systems in heterogeneous infrastructures, where legacy hardware may constrain feedback responsiveness [18, 25, 26]. In materials informatics, this manifests as interpretive mismatches between data-driven models and resource realities, particularly in high-throughput settings where rapid iterations clash with fixed allocations [3, 4, 21]. The layered structure of AREM suggests pathways for mitigation, such as modular upgrades that align with discovery steering logics, but it also highlights the need for interoperability standards to facilitate seamless layer interactions [10, 15, 27].

A critical aspect is the interaction between representation learning and resource orchestration. Deep architectures, including those for crystal and molecular graphs [10, 12, 14], often require specialized environments, and AREM interprets this as an opportunity for adaptive provisioning that enhances representation-inference synergies [6, 8]. However, in multimodal contexts [9, 11], data fusion can introduce overheads that disequilibrate the system, implying a need for refined steering to prioritize high-fidelity modalities without inflating costs [7, 20]. This discussion extends to autonomous discovery, where closed-loop systems [4, 13] rely on real-time feedbacks; AREM's equilibrium model interprets potential instabilities if epistemic signals are delayed, as in simulation-experiment decoupling [12, 18].

Uncertainty quantification emerges as a pivotal interpretive element. By propagating uncertainties through layers [16, 17], AREM fosters robust workflows, but it also reveals systemic vulnerabilities in inverse design [20, 22], where generative sampling might overconsume resources if not equilibrated. Foundation models amplify this, as their generalization capabilities [11, 24] demand balanced pretraining, and the framework interprets trade-offs in transfer learning applications across materials classes [1, 2, 23].

Collaborative implications are noteworthy. In shared ecosystems [19, 27], AREM interprets resource disparities as epistemic risks, suggesting steering logics that promote equitable distribution [28, 29]. Yet, this raises questions about governance, as autonomous agents [4, 13] could inadvertently bias allocations toward dominant datasets [5, 6]. The framework's feedback mechanisms offer a conceptual counter, interpreting iterative adjustments as means to achieve inclusivity in discovery pipelines [10, 21].

Overall, while AREM provides a cohesive lens for infrastructure dynamics, its interpretive nature underscores the absence of empirical benchmarks, positioning it as a tool for conceptual refinement rather than operational prescription. This aligns with evolving paradigms in computational materials science, where scalability hinges on harmonious system interactions [3, 25, 26]. Future conceptual extensions could explore hybrid human-AI steering, further enriching the discourse on resilient discovery infrastructures.

Conclusion

In synthesizing the conceptual underpinnings of scaling discovery infrastructures in computational and data-driven materials engineering, this manuscript has introduced the Adaptive Resource Equilibrium Model (AREM) as a novel framework for interpreting resource allocation dynamics. By structuring ecosystems into interconnected layers with feedback loops, AREM elucidates the balances required to navigate data complexity, model demands, and discovery throughput, drawing from key advancements in materials informatics, machine learning architectures, and autonomous systems.

The analytical implications and discussions highlight AREM's potential to reinterpret trade-offs and epistemic structures, offering systems-level insights that could inform more resilient workflows without empirical assertion. Ultimately, this work contributes a theoretical perspective

on optimizing computational ecosystems, emphasizing adaptive equilibria as central to advancing scalable, efficient materials discovery.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 08 Mar 2024 Revised: 05 Apr 2024 Accepted: 05 May 2024
Published online: 18 September 2024

References

Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C. Machine learning in materials informatics: Recent applications and prospects. npj Comput Mater. 2017;3(1):54.

Pilania G, Gubernatis JE, Lookman T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. Comput Mater Sci. 2017;129:156-63.

Xue H, Cheng G, Yin WJ. Computational design of energy-related materials: From first-principles calculations to machine learning. Wiley Interdiscip Rev Comput Mol Sci. 2024;14(5):e1732.

Kaufmann K, Maryanovsky D, Mellor WM, Zhu C, Rosengarten AS, Vecchio KS. Discovery of high-entropy ceramics via closed-loop optimization using artificial intelligence. npj Comput Mater. 2020;6(1):42.

Schmidt J, Shi J, Borlido P, Chen L, Botti S, Marques MAL. Predicting the thermodynamic stability of solids combining density functional theory and machine learning. Chem Mater. 2017;29(12):5090-103.

Chen C, Zuo Y, Ye W, Li X, Deng Z, Ong SP. A critical review of machine learning of energy materials. Adv Energy Mater. 2020;10(8):1903242.

Kim C, Batra R, Chen L, Tran H, Ramprasad R. Polymer design using genetic algorithm and machine learning. Comput Mater Sci. 2021;186:110067.

Tran H, Mannodi-Kanakkithodi A, Kim C, Sharma V, Pilania G, Ramprasad R. Machine-learning predictions of polymer properties with Polymer Genome. J Appl Phys. 2020;128(17):171104.

Sun S, Hartono NTP, Ren ZD, Oviedo F, Buscemi AM, Layurova M, et al. Accelerated development of perovskite-inspired materials via high-throughput synthesis and machine-learning diagnosis. Joule. 2019;3(6):1437-51.

Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. Chem Mater. 2019;31(9):3564-72.

Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. Nature. 2023;624:80-5.

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.

Kusne AG, Yu H, Wu C, Zhang H, Hattrick-Simpers J, DeCost B, et al. On-the-fly closed-loop materials discovery via Bayesian active learning. Nat Commun. 2020;11(1):5966.

Choudhary K, Kalish I, Beams R, Tavazza F. High-throughput assessment of vacancy formation and surface energies of materials using graph neural networks. Phys Rev Mater. 2021;5(1):013803.

Fung V, Zhang J, Juarez E, Sumpter BG. Benchmarking graph neural networks for materials chemistry. npj Comput Mater. 2021;7(1):84.

Fulginiti D, Abraham Y, Florentin L, Bernadou S. Machine learning-driven discovery of key descriptors for CO2 activation on bimetallic transition metal carbides. Digit Discov. 2022;1(3):320-8.

Saidi WA, Castelli IE, Wisesa P, Jain A. Autonomous materials discovery driven by Gaussian process regression with inhomogeneous measurement noise and anisotropic kernels. Sci Rep. 2020;10(1):17059.

Liu C, Zhang Y, Zhang T, Wu X, Gao L, Zhang Q. High throughput vehicle coordination strategies at road intersections. IEEE Trans Veh Technol. 2020;69(12):14341-54.

Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD). JOM 2013;65(11):1501-9.

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 design of inorganic materials. npj Comput Mater. 2020;6(1):84.

Korolev V, Mitrofanov A, Eliseev A, Tkachenko V. Machine-learning-assisted search for functional materials over extended chemical space. Mater Horiz. 2020;7(10):2710-8.

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

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

Saidi WA, Shadid W, Vesely EJ. Machine-Learning structural and electronic properties of high-temperature superconductors. npj Comput Mater. 2020;6(1):197.

Umeda Y, Hayashi H, Moriwake H, Tanaka I. Prediction of dielectric constants using a combination of first principles calculations and machine learning. Jpn J Appl Phys. 2019;58(SL):SLLC01.

Ping J, Fan Z, Sindoro M, Ying Y, Zhang H. Recent advances in sensing applications of two-dimensional transition metal dichalcogenide nanosheets and their composites. Adv Funct Mater. 2017;27(19):1605817.

Aykol M, Hegde VI, Hung L, Suram SK, Herring P, Wolverton C, et al. Network analysis of synthesizable materials data. Nat Commun. 2019;10(1):2014.

Badini S, Regondi S, Pugliese R. Unleashing the power of artificial intelligence in materials design. Materials. 2023;16(17):5927.

Oliynyk AO, Mar A. Discovery of intermetallic compounds from traditional to machine-learning approaches. Acc Chem Res. 2018;51(1):59-68.