**ORIGINAL RESEARCH** **Open access**

# The Interpretability–Complexity Paradox in Deep Materials Networks

Bruno Martins[1*], Lucas Pereira[1], Renata Azevedo[2]

## Abstract

In the evolving landscape of computational and data-driven materials engineering, deep neural networks have emerged as powerful tools for accelerating materials discovery and design. These architectures leverage vast multimodal datasets, high-throughput computations, and representation learning to model complex structure-property relationships in materials systems. However, a fundamental tension arises: as network complexity increases to capture intricate physical phenomena, interpretability diminishes, hindering the extraction of scientific insights essential for advancing materials informatics. This interpretability-complexity paradox poses a significant barrier to integrating deep models into autonomous discovery pipelines, where uncertainty quantification and simulation-experiment coupling demand transparent decision-making. To address this gap, we introduce the Interpretive Complexity Equilibrium Framework (ICEFrame), a novel conceptual structure that conceptualizes the dynamic interplay between model depth, representational fidelity, and epistemic transparency in deep materials networks. ICEFrame delineates layered interactions across data ingestion, architectural scaling, and inference steering, incorporating feedback mechanisms to balance trade-offs without empirical validation. This framework offers interpretive lenses for navigating complexity in graph neural networks and foundation models for science, fostering more robust closed-loop experimentation and inverse design strategies. By reframing the paradox through systems-level insights, ICEFrame implications extend to enhancing discovery steering logics in materials AI, ultimately promoting sustainable innovation in computational materials ecosystems.

**Keywords** Autonomous discovery, Materials informatics, Interpretability, Graph neural networks, Deep learning, Model complexity

*Correspondence:
Bruno Martins
bruno.martins@gmail.com
[1] Department of Materials Modeling and Data Science, Faculty of Engineering, University of Minho, Braga, Portugal
[2] Department of Computational Materials Systems, Faculty of Engineering, University of Porto, Porto, Portugal

# Introduction

## The emergence of data-driven paradigms in materials engineering

The field of materials engineering has undergone a profound transformation with the advent of data-driven approaches, shifting from traditional trial-and-error methodologies to computationally intensive, predictive frameworks [1, 2]. At the core of this shift lies materials informatics, which integrates machine learning techniques with high-throughput computational tools to expedite the exploration of vast materials spaces [3]. This paradigm leverages multimodal datasets encompassing structural, chemical, and physical properties, enabling the identification of novel materials for applications in energy storage, catalysis, and advanced manufacturing [4]. Deep learning architectures, particularly graph neural networks, have proven instrumental in representing atomic-scale interactions and predicting emergent behaviors in complex systems such as alloys and polymers [5, 6].

The integration of these methods into materials research ecosystems has been facilitated by advancements in data curation and sharing platforms, which aggregate

experimental and simulated data to train robust models [7]. For instance, representation learning strategies have evolved to encode materials' hierarchical features, from atomic graphs to macroscopic properties, enhancing the fidelity of predictions [8]. Yet, as these data-driven pipelines mature, they introduce new challenges in managing the scale and heterogeneity of information flows, necessitating sophisticated computational infrastructures [9].

## The imperative of deep networks in capturing materials complexity

Deep neural networks are indispensable for addressing the inherent complexity of materials systems, where non-linear interactions and multi-scale phenomena defy shallow modeling approaches [10, 11]. In materials science, complexity manifests in the form of diverse phase spaces, defect dynamics, and environmental responses, requiring architectures with multiple layers to approximate underlying physical laws [2]. Graph-based models, for example, excel in handling topological variances in crystal structures, allowing for transferable predictions across material classes [5, 12].

This depth enables inverse design workflows, where desired properties guide the generation of candidate structures, accelerating discovery cycles [13]. Moreover, coupling with high-throughput simulations permits the exploration of hypothetical materials, bridging gaps between theory and experimentation [14]. However, the push for deeper networks amplifies computational demands, prompting innovations in efficient training regimes and uncertainty-aware inferences [15, 16]. Despite these advances, the black-box nature of deep models often obscures the pathways from input representations to output predictions, complicating their adoption in safety-critical materials applications [17].

## Challenges in interpretability within materials AI

Interpretability in machine learning refers to the ability to understand and explain model decisions in human-comprehensible terms, a critical requirement in scientific domains like materials engineering [18, 19]. In deep networks applied to materials, interpretability facilitates the validation of predictions against physical principles, aiding in the refinement of discovery pipelines [20]. Techniques such as feature attribution and surrogate modeling have been employed to dissect network behaviors, revealing how specific inputs influence property forecasts [21, 22].

Nevertheless, achieving interpretability is fraught with obstacles, particularly in data-sparse regimes common to materials datasets [3]. Multimodal integrations, while enriching representations, can introduce confounding factors that obscure causal links [23]. Furthermore, in autonomous systems, where models steer experimental feedbacks, a lack of transparency risks propagating epistemic uncertainties, undermining trust in AI-assisted discoveries [24, 25].

## Delineating the interpretability-complexity paradox

The interpretability-complexity paradox emerges as a central tension in deep materials networks: escalating architectural sophistication enhances predictive power but erodes mechanistic transparency [26, 27]. This paradox is exacerbated in computational materials ecosystems, where models must navigate high-dimensional spaces while providing actionable insights for closed-loop experimentation [28, 29]. On one hand, complexity allows for nuanced capturing of quantum-mechanical effects and thermodynamic stabilities; on the other, it impedes the deconvolution of learned representations into interpretable knowledge [30, 31].

Existing literature highlights this trade-off, with studies noting diminished explainability in scaled-up networks despite improved generalization [4, 6]. The paradox not only affects model deployment but also influences the epistemic foundations of materials discovery, where opaque decisions can lead to overlooked biases or unphysical extrapolations [7, 9]. Addressing this requires a conceptual reevaluation of how complexity and interpretability interact within data-driven workflows.

In light of these considerations, this manuscript positions the Interpretive Complexity Equilibrium Framework (ICEFrame) as a novel lens for understanding and navigating the paradox. ICEFrame integrates systems-level dynamics to conceptualize balanced interactions, paving the way for enhanced computational steering in materials AI.

# Theoretical Background & Literature Synthesis

## Foundations of computational materials ecosystems

Computational materials engineering has undergone a profound epistemic transformation, evolving from a simulation-support discipline into a fully data-centric scientific ecosystem. This transition has been catalyzed by the convergence of high-performance computing, digital infrastructures, and materials informatics frameworks that formalize how materials knowledge is generated, structured, and operationalized [1, 2]. Within this paradigm, data are no longer passive outputs of computational workflows but active substrates through which discovery logics are encoded and enacted.

At the infrastructural core of this ecosystem lie high-throughput computational methodologies, particularly density functional theory (DFT) and large-scale molecular dynamics simulations, which collectively generate expansive, standardized datasets spanning compositional, structural, thermodynamic, and electronic property spaces [10, 14]. These infrastructures—often embedded within federated repositories and open databases—enable systematic traversal of materials design spaces at scales previously unattainable through experimental means alone. As such, they function not merely as data generators but as epistemic accelerators that reshape how hypotheses are formulated and validated.

The proliferation of such data ecosystems has facilitated a decisive shift from descriptive modeling—focused on post-hoc rationalization of observed phenomena—toward predictive and prescriptive analytics capable of guiding materials design ex ante [3, 13]. Machine learning systems trained on curated computational datasets can now forecast properties, identify stability regimes, and propose candidate materials with unprecedented speed, thereby redefining the tempo of scientific iteration.

Representation learning constitutes a critical intermediary layer within this ecosystem, transforming raw atomistic and microstructural configurations into machine-interpretable embeddings [5, 8]. Through learned latent spaces, these representations encode physicochemical relationships that would otherwise remain computationally intractable. Graph neural networks (GNNs), in particular, have emerged as dominant architectures due to their capacity to preserve relational inductive biases inherent in materials systems, including atomic connectivity, bonding topology, and symmetry constraints [6, 12]. Their deployment has enabled breakthroughs in property prediction, phase identification, and defect characterization across crystalline and amorphous domains.

Despite these advances, the literature underscores a structural dependency on standardized benchmarking datasets to ensure cross-model comparability and reproducibility [7, 9]. Without such harmonization, representation learning risks fragmentation, where model performance becomes contingent on dataset idiosyncrasies rather than scientific validity. Consequently, dataset curation, annotation fidelity, and ontological alignment have become central governance concerns within computational materials ecosystems.

## Advances in deep learning architectures for materials science

Deep learning has introduced a new architectural paradigm capable of accommodating the multiscale and multimodal complexity intrinsic to materials phenomena. Unlike traditional machine learning pipelines reliant on handcrafted descriptors, deep architectures autonomously extract hierarchical features directly from raw or minimally processed inputs [2, 11]. This capacity aligns closely with the scale-bridging nature of materials science, where macroscopic properties emerge from nested interactions spanning electronic, atomic, and microstructural levels.

Convolutional neural networks (CNNs) have demonstrated efficacy in processing spatially resolved materials data, including microstructural imaging and spectroscopic mappings, enabling automated defect detection and morphology classification [15, 26]. Attention-based and transformer architectures extend this capability by dynamically weighting feature contributions, allowing models to prioritize salient structural motifs or compositional interactions within high-dimensional inputs.

For crystalline and polymeric systems, architectural innovations increasingly incorporate symmetry-aware operations that embed physical invariances—such as rotational, translational, and permutational symmetries—directly into network design [8, 17]. These physics-aligned inductive biases enhance generalization across compositional families and structural phases, mitigating overfitting while preserving mechanistic coherence.

The literature also documents the emergence of scientific foundation models—large-scale architectures pre-trained on heterogeneous scientific corpora and subsequently fine-tuned for domain-specific materials tasks [19, 29]. By integrating multimodal inputs—including spectroscopy, microscopy, simulation outputs, and textual literature—these models enable cross-modal reasoning and knowledge transfer, expanding the functional scope of autonomous discovery platforms [23, 24].

However, the scaling of deep architectures introduces non-trivial computational overheads, including increased training costs, energy consumption, and parameter redundancy [20, 30]. These burdens have stimulated parallel research into efficient model variants—such as sparse networks, distillation frameworks, and low-rank approximations—that seek to preserve predictive performance while reducing infrastructural strain. Thus, architectural innovation in materials AI is increasingly defined not only by capability expansion but also by computational sustainability considerations.

## Interpretability techniques in data-driven materials research

As deep learning architectures grow in representational capacity, interpretability has emerged as a critical epistemic requirement rather than an optional analytical add-on. In materials science, predictive accuracy alone is insufficient; models must also facilitate mechanistic insight to support theory development and rational design.

Post-hoc interpretability techniques have therefore gained traction as tools for interrogating learned representations. Saliency mapping approaches visualize feature importance across input spaces, revealing which atomic environments, bonding motifs, or microstructural features most strongly influence model outputs [18, 21, 22]. Concept activation vector methodologies extend this paradigm by mapping latent features to human-interpretable scientific constructs, such as crystallographic symmetries or defect typologies.

Applications of these techniques have yielded substantive insights into structure–property relationships, particularly in alloy optimization and composite design, where identifying critical compositional drivers can accelerate targeted experimentation [4, 27]. By translating abstract latent encodings into domain-meaningful descriptors, interpretability tools bridge the gap between algorithmic inference and scientific reasoning.

Uncertainty quantification (UQ) operates as a complementary interpretive layer, providing probabilistic confidence estimates that contextualize predictive outputs [16, 25, 28]. Bayesian neural networks, ensemble learning, and variational inference frameworks enable decomposition of epistemic versus aleatoric uncertainties, informing risk-aware decision-making in materials screening and deployment [3, 31].

Yet interpretability remains structurally constrained by the opacity of high-depth architectures. Inverse design systems—where models generate candidate materials rather than evaluate them—exemplify this limitation: outputs often lack transparent causal rationales, complicating downstream validation and synthesis planning [13, 14]. Consequently, interpretability research increasingly advocates for intrinsically interpretable architectures rather than reliance on retrospective explanatory overlays.

## Uncertainty and feedback in autonomous discovery pipelines

Autonomous discovery platforms represent the operational apex of computational materials ecosystems, integrating AI prediction engines with robotic synthesis and characterization systems in closed-loop configurations [7, 29]. Within these pipelines, machine learning models iteratively guide experimental selection, creating self-refining discovery cycles that compress traditional research timelines.

Active learning strategies constitute the decision-theoretic backbone of such systems. By prioritizing experiments that maximize expected information gain or minimize predictive uncertainty, these algorithms optimize resource allocation across vast candidate spaces [9, 24]. Feedback loops generated through iterative simulation–experiment coupling enable continuous recalibration of predictive models, enhancing both accuracy and coverage.

Uncertainty metrics function as steering signals within these loops, directing exploration toward underrepresented or epistemically volatile regions of materials space [15, 16]. This adaptive sampling logic not only accelerates discovery but also mitigates dataset bias accumulation over successive experimental cycles.

However, multimodal uncertainty integration remains an unresolved systems challenge. Discrepancies between simulated predictions and experimental measurements—

arising from model approximations, synthesis imperfections, or measurement noise—can propagate through closed loops, amplifying epistemic error [23, 28]. Distributional shifts further destabilize predictive reliability, particularly when autonomous systems encounter materials regimes absent from training corpora [2, 20].

These dynamics necessitate robust calibration frameworks capable of harmonizing heterogeneous uncertainty sources while preserving loop stability.

## Gaps in addressing the interpretability–complexity trade-off

A cross-sectional synthesis of the literature reveals a persistent structural tension between interpretability and architectural complexity in deep materials learning systems [18, 26, 30]. Graph-based and multimodal architectures have significantly enhanced representational fidelity, enabling nuanced modeling of physicochemical interactions. Yet this complexity often comes at the expense of transparency, producing predictive systems that function as epistemic black boxes [5, 6, 12].

Existing interpretability methods, while methodologically sophisticated, are frequently retrofitted onto pre-existing architectures rather than co-designed within them [4, 21, 22]. This retrospective orientation limits their explanatory granularity and risks producing superficial rationalizations rather than mechanistic insight.

Within autonomous discovery ecosystems, this opacity introduces operational risks. Steering decisions—such as candidate prioritization or experimental pathway selection—become difficult to audit when underlying inference logics remain inscrutable [7, 25, 29]. Consequently, interpretability deficits can propagate from model design into experimental governance, constraining trust and reproducibility.

Uncertainty quantification partially mitigates these risks by flagging low-confidence predictions, yet its interaction with architectural complexity remains conceptually underdeveloped [16, 28, 31]. Most implementations treat uncertainty as an auxiliary output rather than an integrated structural feature of model reasoning.

The literature therefore calls for integrative conceptual frameworks that jointly theorize representation complexity, interpretability infrastructures, and uncertainty propagation within unified systems architectures [1, 2, 10]. Such frameworks would transcend incremental methodological refinements, offering epistemic design principles for balancing predictive power with scientific intelligibility.

This synthesis foregrounds the necessity of interpretive systems thinking in computational materials engineering—where discovery acceleration must be harmonized with explanatory depth to ensure epistemic reliability, governance accountability, and sustainable innovation trajectories [3, 11, 17].

## Proposed conceptual framework

The Interpretive Complexity Equilibrium Framework (ICEFrame) is introduced as an original conceptual structure to navigate the interpretability-complexity paradox in deep materials networks. ICEFrame posits a layered systems architecture that delineates the flow from data representations to discovery outcomes, emphasizing dynamic equilibria rather than static optimizations. At its core, ICEFrame comprises four interconnected layers: the Data Representation Layer, the Architectural Scaling Layer, the Inference Steering Layer, and the Discovery Integration Layer. These layers facilitate a pipeline where raw multimodal materials data—encompassing atomic structures, property spectra, and simulation outputs—are transformed into interpretable insights through iterative feedback loops.

In the Data Representation Layer, inputs are encoded via graph-based or vectorial schemes, setting the foundation for subsequent complexity buildup [5, 8]. This transitions to the Architectural Scaling Layer, where network depth and width modulate to capture intricate interactions, such as electronic band structures or mechanical responses [2, 11]. Here, complexity is conceptualized as a scalable parameter influencing representational capacity.

The Inference Steering Layer introduces computational logics that guide model behaviors, incorporating uncertainty signals to adjust transparency levels [16, 28]. Finally, the Discovery Integration Layer synthesizes outputs into actionable knowledge, feeding back to earlier layers to refine representations and architectures [7, 29].

Feedback loops in ICEFrame operate bidirectionally: an interpretability feedback loop propagates transparency constraints upward from inference to architecture, while a complexity feedback loop cascades depth requirements downward from discovery needs. These loops enable

equilibrium states where trade-offs are dynamically balanced, enhancing workflow resilience in autonomous systems.

The interplay between interpretability (I) and complexity (C) can be conceptualized as a relational dynamic $: I \approx \frac{\kappa}{(C^{\alpha} + \beta \cdot U)}$, where κ represents a domain-specific transparency constant, α denotes the scaling exponent of architectural depth, β weights uncertainty contributions (U), and the approximation captures the inverse relationship tempered by epistemic factors. This expression may be expressed as a guiding heuristic for steering logics, illustrating how escalating C diminishes I unless mitigated by U-aware adjustments.

Furthermore, the feedback loop dynamics capture the interaction between layers as $\Delta L_n = \gamma \cdot \left(I_{\{n-1\}} - \theta \cdot C_n\right)$, where ΔL_n is the adjustment to layer n, γ is a convergence factor, θ a threshold for complexity tolerance, and subscripts denote layer transitions. This formalizes the iterative equilibration, ensuring systems-level coherence.

Representation-inference interactions are further modeled as $R \otimes Inf = \Sigma \varphi_i \cdot w_i$, where R is the representation tensor, Inf the inference operator, $\varphi_i$ feature mappings, and w_i weights reflecting paradox-induced distortions. This captures how complex representations entangle with opaque inferences, with ICEFrame's loops disentangling them through steering.

As conceptualized in the layered systems architecture of ICEFrame, interpretability and complexity interact through bidirectional feedback equilibria across discovery pipelines **(Figure 1)**.

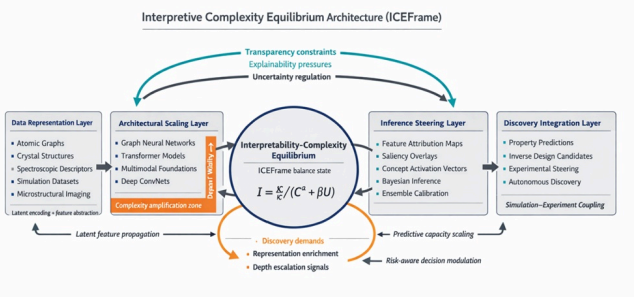


**Figure 1**. The Interpretive Complexity Equilibrium Framework (ICEFrame)

**Figure 1** illustrates the Interpretive Complexity Equilibrium Framework (ICEFrame), a layered conceptual architecture that models the interpretability–complexity paradox in deep materials learning systems. The framework comprises four pipeline strata—Data Representation, Architectural Scaling, Inference Steering, and Discovery Integration—through which multimodal materials data are transformed into actionable discovery outputs. Increasing architectural depth amplifies representational fidelity but introduces interpretability attenuation, visualized as complexity pressure zones. A central equilibrium node conceptualizes the dynamic balance between transparency and predictive capacity. Bidirectional feedback loops regulate this balance: interpretability constraints propagate upward from inference layers, while discovery demands drive complexity scaling downward. Uncertainty quantification operates as a mediating channel, steering inference calibration and experimental prioritization. ICEFrame thus provides a systems-level lens for analyzing epistemic trade-offs in autonomous materials discovery infrastructures.

## Analytical implications

### Systems-level insights into discovery pipelines

The ICEFrame provides a lens for analyzing how the interpretability-complexity paradox influences data-to-discovery pipelines in computational materials engineering [1, 2]. By conceptualizing layers as interconnected modules, ICEFrame highlights the propagation of trade-offs across workflows, where architectural scaling in deep networks can amplify epistemic risks if not counterbalanced by steering logics [2, 11]. For instance, in high-throughput computation scenarios, enhanced complexity enables finer-grained property predictions, yet it may obscure the mapping from multimodal inputs to outputs, complicating the integration of simulation data with experimental validations [14, 23].

This implies a need for dynamic adjustments in representation learning, where graph neural networks' depth is modulated to preserve transparency in structure-property inferences [5, 6, 8]. Analytically, ICEFrame suggests that feedback loops can mitigate paradox-induced distortions by iteratively refining layer interactions, fostering resilient pipelines that adapt to data heterogeneity [7, 9]. Such insights extend to autonomous discovery systems, where closed-loop feedbacks rely on transparent

inferences to steer towards optimal materials candidates [7, 29]. The layered propagation of interpretability and complexity trade-offs across discovery infrastructures can be systematically synthesized through ICEFrame's functional strata **(Table 1)**.

**Table 1.** Layered Trade-Off Structures in the Interpretability–Complexity Equilibrium Framework (ICEFrame)

| ICEFrame Layer | Functional Role | Complexity Drivers | Interpr Mech |
|---|---|---|---|
| Data Representation Layer | Encodes multimodal materials inputs into latent embeddings | Descriptor dimensionality; multimodal fusion; graph connectivity scaling | Fea visual desc redu embe clus |
| Architectural Scaling Layer | Expands representational capacity through deep architectures | Network depth; parameter volume; multimodal integration | Sym cons sp archite phy informe |
| Inference Steering Layer | Governs prediction logic and interpretive extraction | Non-linear inference pathways; attention complexity | Sal map cor vec surr mo |
| Discovery Integration Layer | Translates predictions into experimental and design actions | Closed-loop automation; candidate generation scale | Exper valic mech cross- |
| Feedback Equilibrium Loops | Regulate trade-offs between depth and transparency | Iterative model scaling; adaptive retraining | Expla feed interp cons |

## Epistemic risk structures in materials AI

Epistemic risks in deep materials networks arise from the paradox, manifesting as uncertainties in model generalizations across diverse datasets [16, 25, 28]. ICEFrame's framework interprets these risks through layer-specific vulnerabilities: in the Data Representation Layer, incomplete encodings introduce biases; in the Architectural Scaling Layer, over-parameterization leads to overfitting disguised as sophistication [3, 10]. Analytically, this structures risks as cascading effects, where unmitigated complexity erodes confidence in inverse design outcomes [13, 31].

The uncertainty-complexity interaction can be expressed as $U \approx \lambda \cdot (C - \mu \cdot I)$, where $\lambda$ is a risk amplification factor, $\mu$ a mitigation coefficient linking interpretability to complexity thresholds, and the approximation captures how excess C escalates U unless I intervenes. This formalization implies computational strategies that prioritize equilibrium states, reducing risks in foundation models for science by embedding transparency constraints [19, 24].

Furthermore, in uncertainty quantification contexts, ICEFrame implies enhanced risk assessment by tracing feedback paths, enabling the identification of high-risk inference points in materials informatics workflows [4, 18, 26].

## Representation-inference interactions and trade-offs

ICEFrame elucidates the interplay between representations and inferences, where complex deep architectures transform atomic graphs into predictions but often at the cost of decipherable pathways [12, 22, 30]. Analytically, this interaction reveals trade-offs in multimodal datasets, as enriched representations demand deeper networks, potentially fragmenting inference logics [23, 27]. The framework's loops offer interpretive tools to analyze these dynamics, suggesting that steering mechanisms can realign representations with interpretable inferences, optimizing for knowledge extraction in polymer or alloy systems [8, 17].

Infrastructure trade-offs emerge prominently: computational resources allocated to depth may detract from interpretability modules, implying a need for balanced resource distribution in discovery steering [15, 20]. The trade-off dynamic may be conceptualized as $T = \int (C \cdot dR - I \cdot dInf)$, integrating over representation (R) and inference (Inf) differentials, where T represents the net trade-off cost, capturing the cumulative impact of paradox imbalances.

This analytical perspective fosters insights into how ICEFrame can guide the design of hybrid architectures,

blending complexity with transparency for robust materials AI applications [15, 21].

## Computational workflow dynamics in closed-loop systems

In closed-loop experimentation, ICEFrame's implications center on workflow dynamics, where paradox resolution enhances the coupling of simulations and experiments [14, 29]. Analytically, the framework interprets these dynamics as equilibrium-driven processes, with feedback loops synchronizing layer outputs to minimize epistemic divergences [7, 24]. This implies improved steering logics that adapt network complexity to real-time data feedbacks, accelerating convergence in autonomous discovery [9, 28].

For graph neural networks in phase prediction, ICEFrame suggests analytical pathways to evaluate how complexity scaling affects workflow efficiency, without empirical metrics [5, 12]. Overall, these implications underscore ICEFrame's role in reconfiguring computational infrastructures for paradox-aware materials engineering [1, 10].

# Results and Discussion

The ICEFrame advances conceptual understanding in computational and data-driven materials engineering by reframing the interpretability-complexity paradox through layered systems dynamics [11, 18, 26]. Unlike prior syntheses that focus on technique-specific enhancements, ICEFrame integrates pipeline-wide interactions, offering a holistic view of how depth and transparency coexist in deep networks [2, 22, 30]. This discussion explores the broader ramifications, emphasizing interpretive shifts in materials informatics and AI ecosystems.

One key aspect is the framework's alignment with evolving paradigms in representation learning and graph architectures [5, 6, 8]. By delineating feedback loops, ICEFrame facilitates a nuanced analysis of how multimodal data influences architectural choices, potentially guiding the development of more adaptive models in high-throughput settings [7, 14, 23]. This interpretive approach contrasts with literature emphasizing empirical interpretability add-ons, instead prioritizing inherent equilibrium mechanisms [4, 21].

In uncertainty quantification, ICEFrame's structures provide a basis for interpreting epistemic challenges in foundation models, where complexity often masks underlying variabilities [16, 19, 25]. The framework's formulas conceptualize these as interactive terms, suggesting pathways for embedding uncertainty-aware steering without altering core architectures [28, 31]. This has implications for inverse design, where transparent workflows could enhance the reliability of generated materials candidates [13, 27].

Furthermore, ICEFrame contributes to discourse on autonomous discovery by conceptualizing closed-loop systems as paradox-resilient entities [9, 29]. Feedback dynamics imply a shift towards self-regulating pipelines, where discovery steering logics evolve to balance trade-offs, fostering innovation in simulation-experiment couplings [15, 24]. However, this raises considerations for infrastructure scalability: deeper networks demand computational resources that may constrain accessibility in diverse research environments [10, 20].

The framework also intersects with small-data challenges in materials science, interpreting how limited datasets exacerbate the paradox by forcing reliance on complex transfers [3, 12]. Analytically, ICEFrame's layers suggest strategies for representation optimization, potentially mitigating biases in underrepresented material classes [1, 17].

Critically, while ICEFrame remains conceptual, its systems-level insights invite integration with emerging AI trends, such as hybrid human-AI collaborations in materials design [2, 4]. This could extend to epistemic risk management, where transparent complexities support ethical deployments in critical applications [16, 18].

Overall, ICEFrame enriches the theoretical landscape, prompting reevaluations of how deep materials networks are conceptualized within data-driven ecosystems [2, 11, 22].

# Conclusion

The interpretability-complexity paradox in deep materials networks represents a pivotal challenge in advancing computational and data-driven materials engineering. Through the introduction of the Interpretive Complexity Equilibrium Framework (ICEFrame), this manuscript has conceptualized a novel structure that integrates layered pipelines, feedback loops, and steering logics to navigate this tension. By formalizing key dynamics symbolically,

ICEFrame offers interpretive tools for understanding trade-offs without empirical assertions, emphasizing systems-level equilibria.

Analytical implications derived from ICEFrame illuminate epistemic risk structures, representation-inference interactions, and workflow dynamics, providing lenses for enhancing discovery processes in materials informatics. These insights underscore the framework's potential to foster resilient autonomous systems, where balanced complexities support transparent innovations.

In discussion, ICEFrame's contributions highlight shifts towards adaptive architectures and uncertainty-aware designs, aligning with broader trends in AI for science. Ultimately, this conceptual advancement promotes a more integrative approach to deep learning in materials, paving the way for sustainable progress in the field.

## Acknowledgements

None

## Conflict of interest

None

## Financial support

None

## Ethics statement

None

Received: 03 Mar 2023 Revised: 28 Apr 2023 Accepted: 16 May 2023
Published online: 18 September 2023



## References

Durmaz AR, Müller M, Lei B, Thomas A, Britz D, Holm EA, et al. A deep learning approach for complex microstructure inference. Nat Commun. 2021;12:6272. https://doi.org/10.1038/s41467-021-26565-5.

Schmidt J, Marques MRG, Botti S, Marques MAL. Recent advances and applications of machine learning in solid-state materials science. npj Comput Mater. 2019;5:83. https://doi.org/10.1038/s41524-019-0221-0.

Xu P, Ji X, Li M, Lu W. Small data machine learning in materials science. npj Comput Mater. 2023;9:42. https://doi.org/10.1038/s41524-023-01000-z.

Kailkhura B, Gallagher B, Kim S, Hiszpanski A, Han TY. Reliable and explainable machine-learning methods for accelerated material discovery. npj Comput Mater. 2019;5:108. https://doi.org/10.1038/s41524-019-0248-2.

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

Gupta V, Choudhary K, DeCost B, Tavazza F, Campbell C, Liao WK, et al. Structure-aware graph neural network based deep transfer learning framework for enhanced predictive analytics on diverse materials datasets. npj Comput Mater. 2023;9:185.

Hu J, Stefanov S, Song Y, Omee SS, Louis SY, Siriwardane EMD, et al. MaterialsAtlas.org: A materials informatics web app platform for materials discovery and survey of state-of-the-art. npj Comput Mater. 2022;8:65. https://doi.org/10.1038/s41524-022-00750-6.

Queen O, McCarver GA, Thatigotla S, Abolins BP, Brown CL, Maroulas V, et al. Polymer graph neural networks for multitask property learning. npj Comput Mater. 2023;9:90. https://doi.org/10.1038/s41524-023-01034-3.

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

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

Schmidt J, Shi SQ, Chen R, Ben-Zion I, Da B, Liu S, et al. Recent advances and applications of deep learning methods in materials science. npj Comput Mater. 2022;8:59. https://doi.org/10.1038/s41524-022-00734-6.

Chang R, Wang YX, Ertekin E. Towards overcoming data scarcity in materials science: unifying models and datasets with a mixture of experts framework. npj Comput Mater. 2022;8:242. https://doi.org/10.1038/s41524-022-00929-x.

Hidalgo F, Schmidt J, Botti S, Marques MAL. Interpreting economic complexity. Sci Adv. 2019;5(1):eaau1705. https://doi.org/10.1126/sciadv.aau1705.

Brynjolfsson E, Hui X, Liu M. What can machine learning do? Workforce implications. Science. 2017;358(6370):1530-4. https://doi.org/10.1126/science.aap8062.

Hughes TWP, Williamson M, Grover A, Ermon S. Wave physics as an analog recurrent neural network. Sci Adv. 2019;5(12):eaay6946. https://doi.org/10.1126/sciadv.aay6946.

Vinuesa R, Sirmacek B. Interpretable deep-learning models to help achieve the sustainable development goals. Nat Mach Intell. 2021;3:926. https://doi.org/10.1038/s42256-021-00414-y.

Brown N, Sandholm T. Superhuman AI for multiplayer poker. Science. 2019;365(6456):885-90. https://doi.org/10.1126/science.aay2400.

Li Z, Nguyen TT, Galindo Torres SA, Peille P, Farahani F, et al. Interpretable deep-learning for guided microstructure-property explorations in photovoltaics. npj Comput Mater. 2019;5:231. https://doi.org/10.1038/s41524-019-0231-y.

Lauritsen CL, Østergaard MS, Kongsø JH, Lauritsen L, Jørgensen MJ, Lange J, et al. A cross-species neural integration of gravity for motor optimization. Sci Adv. 2021;7(15):abf7800. https://doi.org/10.1126/sciadv.abf7800.

Bottaro S, Lindorff-Larsen K. Biophysical experiments and biomolecular simulations: A perfect match? Science. 2018;361(6400):355-60. https://doi.org/10.1126/science.aat4010.

Frégnac Y. Big data and the industrialization of neuroscience: A safe roadmap for understanding the brain? Science. 2017;358(6362):470-7. https://doi.org/10.1126/science.aan8866.

Ma B, Wu X, Zhao C, Lin C, Gao M, Sa B, et al. An interpretable machine learning strategy for pursuing high piezoelectric coefficients in (K0.5Na0.5)NbO3-based ceramics. npj Comput Mater. 2023;9:229. https://doi.org/10.1038/s41524-023-01187-1.

Nathan R, Monk CT, Arlinghaus R, Adam T, Alós J, Assaf , et al. Big-data approaches lead to an increased understanding of the ecology of animal movement. Science. 2023;379(6632):eabg1780.

Marcinkevičs R, Ozkan A, Wolski A, Vogelsanger P, Baumann S, Slodczyk B, et al. Interpretable and explainable machine learning: A methods-centric overview with concrete examples.Wiley Interdiscip Rev Data Min Knowl Discov. 2023;13(3):e1493. https://doi.org/10.1002/widm.1493.

Fukui S, Ochi M, Sakurai Y. Interpretable machine learning for maximum corrosion depth and influence factor analysis. npj Mater Degrad. 2023;7:24. https://doi.org/10.1038/s41529-023-00324-x.

Zhu T, Sun W. Explainable machine learning in materials science. npj Comput Mater. 2022;8:184.

Wei J, Chu X, Sun XY, Xu K, Deng HX, Chen J, et al. Machine learning in materials science. InfoMat. 2019;1(3):338-58. https://doi.org/10.1002/inf2.12028.

Yu J, Wang D, Duan C, Liu J, Li Y, Zhou J. Uncertainty quantification: Can we trust artificial intelligence in drug discovery? iScience. 2022;25(8):104814. https://doi.org/10.1016/j.isci.2022.104814.

Feng S, Fu H, Zhou H, Wu Y, Lu Z, Dong H. A general and transferable deep learning framework for predicting phase formation in materials. npj Comput Mater. 2021;7:10. https://doi.org/10.1038/s41524-020-00488-z.

Bhowmik R, Jain M, Ghosh S, Chattopadhyay A. Towards understanding structure–property relations in materials with interpretable deep learning. npj Comput Mater. 2023;9:199.

Li Y, Tang Y, Luo L, Shen B, Zhang F. Infusing theory into deep learning for interpretable reactivity prediction. Nat Commun. 2021;12:5288. https://doi.org/10.1038/s41467-021-25639-8.