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

# Material Spaces Are Not Euclidean: A Computational Critique of Distance Metrics in Data-Driven Materials Discovery

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

## Abstract

In the rapidly evolving field of computational materials engineering, data-driven approaches have transformed the discovery and design of novel materials by leveraging machine learning and high-throughput computations to navigate vast chemical spaces. Traditional methodologies often rely on Euclidean distance metrics to quantify similarities between materials in latent representations, facilitating tasks such as property prediction, inverse design, and autonomous experimentation. However, this assumption overlooks the inherent non-linearities and topological complexities of material spaces, where properties like electronic bandgaps, mechanical strengths, and thermodynamic stabilities emerge from intricate atomic interactions that do not conform to flat geometries. This conceptual gap leads to inefficiencies in representation learning, biased uncertainty quantification, and suboptimal steering in discovery pipelines. Here, we introduce a novel interpretive framework that critiques Euclidean metrics through a manifold-based lens, emphasizing geodesic distances and curvature-aware embeddings to better capture the epistemic structure of materials data. By integrating insights from graph neural networks, multimodal datasets, and closed-loop systems, this framework reveals computational trade-offs in data infrastructures and enhances the interpretability of AI-guided workflows. Implications extend to improved coupling of simulations and experiments, fostering more robust foundation models for materials science and accelerating innovation in energy, electronics, and structural applications without empirical validation.

**Keywords** Materials informatics, Representation learning, Computational discovery, Data-driven design, Machine learning in materials, Non-Euclidean geometries

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

## Introduction

### Euclidean geometries, data-driven design, AI infrastructures

The advent of computational materials engineering has marked a paradigm shift in how scientists and engineers approach the discovery and optimization of new materials. Historically, materials development relied on empirical trial-and-error methods, constrained by the limitations of physical experimentation and human intuition. Knowledge accumulation was incremental, localized, and often bounded by laboratory throughput and instrumentation access. However, with the integration of advanced computational tools, including density functional theory (DFT) and molecular dynamics simulations, the field has transitioned toward predictive modeling capable of exploring expansive parameter spaces [1, 2]. These simulation paradigms enabled the systematic probing of thermodynamic stability, electronic structures, and kinetic

pathways prior to synthesis, effectively repositioning computation as a pre-experimental discovery engine.

This evolution is particularly evident in the rise of high-throughput computational frameworks, which enable the systematic screening of thousands to millions of candidate materials for targeted properties such as superconductivity, catalytic activity, or photovoltaic efficiency [3, 4]. High-throughput infrastructures transform discovery from isolated hypothesis testing into combinatorial exploration, where algorithmically generated libraries are evaluated across multi-property optimization landscapes. The scale of this computational acceleration has redefined feasibility boundaries, allowing exploration of chemical subspaces that would be experimentally inaccessible within conventional time horizons.

Central to this transformation is the role of artificial intelligence (AI) and data ecosystems, which have become indispensable for handling the deluge of information generated by these computations. Materials informatics, a subdiscipline that applies data science principles to materials research, has emerged as a cornerstone, utilizing machine learning algorithms to extract patterns from large datasets [5, 6]. Within this paradigm, AI does not merely automate analysis but actively constructs predictive knowledge structures—mapping latent correlations between composition, structure, processing, and performance.

For instance, supervised learning models predict material properties from compositional and structural descriptors, while unsupervised techniques cluster materials based on similarity measures [7, 8]. These clustering regimes often function as proto-taxonomies of materials space, organizing compounds into property-relevant families that guide subsequent exploration. Dimensionality reduction further enables visualization of latent discovery terrains, providing interpretive windows into high-dimensional chemical complexity.

The proliferation of open-access databases, such as the Materials Project and Automatic FLOW for Materials Discovery (AFLOW), has further fueled this data-driven paradigm, providing standardized repositories that facilitate collaborative research and model training [9, 10]. These infrastructures institutionalize data reproducibility, metadata harmonization, and interoperable benchmarking, thereby establishing the epistemic scaffolding upon which AI-driven discovery operates.

High-throughput infrastructures exemplify the synergy between computation and data analytics, automating workflows from structure generation to property evaluation [11, 12]. Pipeline orchestration platforms integrate simulation engines, feature extraction modules, and predictive models into continuous evaluation loops. These systems often incorporate active learning strategies, where AI models iteratively select the most informative candidates for further simulation or experimentation, thereby optimizing resource allocation [13, 14]. Selection policies are guided by acquisition functions balancing uncertainty, diversity, and performance thresholds.

In autonomous discovery systems, this is extended through closed-loop setups that couple computational predictions with robotic synthesis and characterization, reducing human intervention and accelerating iteration cycles [15, 16]. Here, AI evolves from analytical assistant to experimental orchestrator—steering laboratory action through predictive prioritization.

Yet, despite these advances, current discovery models face significant limitations rooted in their underlying assumptions about material spaces.

## Euclidean constraint in materials representation

One prominent constraint is the epistemic challenge of representing complex materials phenomena within simplified mathematical frameworks. Materials properties arise from quantum-mechanical interactions across multiple scales—from atomic orbitals to macroscopic behaviors—creating datasets that are inherently sparse, noisy, and high-dimensional [17, 18]. These multiscale dependencies generate relational structures that resist linear encoding.

Traditional approaches often project these into Euclidean spaces for ease of computation, employing metrics like cosine similarity or L2 norms to gauge distances between representations [19, 20]. Euclidean embeddings enable tractable optimization but impose geometric assumptions of flatness, isotropy, and linear separability.

However, this overlooks the non-Euclidean nature of material manifolds, where distances may follow curved paths due to hierarchical dependencies, phase transitions, or compositional discontinuities [21, 22]. Phase boundaries, defect topologies, and metastable basins introduce

curvature gradients that distort straight-line similarity metrics.

Such mismatches can distort inference processes, leading to erroneous predictions in regions of sparse data or extrapolation beyond trained domains [23, 24]. **(Table 1)** Materials positioned near manifold edges—such as novel alloys or low-symmetry compounds—are particularly vulnerable to predictive instability.

**Table 1.** Euclidean vs Manifold-Aware Paradigms in Data-Driven Materials Discovery

| Dimension | Euclidean Geometry Paradigm | Manifold-Aware Perspective | Discov Implicat |
|---|---|---|---|
| Distance Metric | Straight-line (L2, cosine) similarity | Geodesic distance along curved manifolds | More physica faithfu similari mappir |
| Materials Space Structure | Flat, linearly separable embedding | Curved, topology-dependent landscape | Preserv phase boundar and discontinu |
| Data Sparsity Interpretation | Treated as absence or noise | Treated as curvature expansion | Protec explorat chemic region |
| Representation Learning | Euclidean latent embeddings | Hyperbolic / spherical embeddings | Captur hierarch compositi diversi |
| GNN Message Passing | Linear neighborhood aggregation | Geodesic relational propagation | Enhanc long-rar depende modeli |
| Inverse Design Navigation | Linear interpolation trajectories | Curvature-aligned optimization paths | Improv feasibilit generat candida |
| Multimodal Data Fusion | Metric homogenization | Geometry-sensitive alignment | Preserv modalit specifi structu |
| Uncertainty Quantification | Gaussian variance fields | Curvature-dependent epistemic risk | Localiz predicti instabil |
| Infrastructure Query Logic | Distance-based retrieval | Topology-aware search indexing | Improv discove targetir |
| Epistemic Risk Profile | Overconfidence in sparse zones | Distortion-aware uncertainty mapping | Reduc extrapola bias |

## Architectural and infrastructural implications

Computationally, these limitations manifest in inefficiencies within AI architectures, particularly graph neural networks (GNNs) tailored for molecular and crystalline structures [25, 26]. While GNNs excel at capturing local symmetries and invariances, their reliance on Euclidean embeddings can amplify errors in global topology, affecting tasks like inverse design where generating viable candidates requires accurate navigation of latent spaces [16, 27].

Uncertainty quantification, essential for reliable decision-making in discovery pipelines, is similarly compromised, as standard probabilistic models assume Gaussian distributions incompatible with manifold curvatures [14, 28]. Predictive confidence may therefore reflect geometric density rather than epistemic validity.

Moreover, the integration of multimodal data—combining simulation outputs, experimental spectra, and literature-mined insights—exacerbates these issues, as disparate modalities introduce heterogeneous metrics that resist uniform Euclidean treatment [22, 29]. Alignment across modalities requires curvature-sensitive fusion strategies capable of preserving relational structure.

Foundation models for science, inspired by large language models, aim to unify these through pre-training on vast corpora, but their effectiveness hinges on robust representation schemes that accommodate non-linear geometries [8, 17]. Without geometric adaptability, large-scale pretraining risks amplifying structural distortions embedded in source datasets.

The epistemic risks here include overconfidence in model outputs and overlooked biases in data curation, which can steer discovery toward local optima rather than global innovations [6, 30].

## Positioning the conceptual intervention

In light of these constraints, there is a pressing need for a reevaluation of distance metrics in data-driven materials discovery. This manuscript positions a new conceptual framework that critiques the Euclidean paradigm, advocating for manifold-aware approaches to enhance computational workflows. By interpreting material spaces through curvature and geodesic lenses, we aim to uncover systems-level insights that refine representation learning, improve pipeline dynamics, and mitigate epistemic trade-offs, ultimately advancing the infrastructure of materials AI.

# Theoretical Background & Literature Synthesis

## Materials data infrastructures

The foundation of data-driven materials discovery lies in robust data infrastructures that aggregate, standardize, and disseminate information across computational and experimental domains. These ecosystems have evolved to support high-throughput paradigms, where vast libraries of material properties are generated via automated simulations [1, 9]. Beyond mere storage repositories, contemporary infrastructures function as epistemic coordination systems—aligning density functional theory outputs, combinatorial synthesis records, and literature-derived descriptors into interoperable discovery substrates.

Key challenges include data sparsity and heterogeneity, as materials datasets often span disparate sources like DFT calculations, experimental measurements, and text-mined literature [10, 22]. This heterogeneity is not only structural but ontological: simulation data encode idealized thermodynamic states, whereas experimental records embed synthesis conditions, defects, and processing histories. Integrating these modalities requires infrastructures capable of reconciling fundamentally different uncertainty regimes and measurement fidelities.

Tools such as Matminer have been developed to facilitate data mining, enabling feature extraction and preprocessing tailored to materials-specific descriptors [10]. These tools operationalize domain-aware featurization—capturing electronegativity gradients, orbital interactions, and crystallographic symmetries. However, the underlying assumption of Euclidean compatibility in these infrastructures can lead to distorted aggregations, where distance-based queries fail to capture topological nuances in chemical spaces [7, 19]. As a result, structurally distant materials may appear artificially proximal, particularly when phase discontinuities or metastability regions are flattened during preprocessing.

This affects the scalability of multimodal datasets, which integrate structural, electronic, and thermodynamic data, requiring infrastructures that accommodate non-linear alignments for effective querying and retrieval [8, 20]. Without curvature-sensitive indexing, discovery platforms risk privileging densely sampled compositional clusters while underrepresenting exploratory chemical frontiers.

## Representation learning architectures

Representation learning forms the computational backbone for encoding materials into machine-readable formats, with architectures like graph neural networks dominating due to their ability to model atomic connectivity [6, 26]. By treating atoms as nodes and bonds as edges, these architectures preserve relational structure across periodic and non-periodic systems.

These models learn embeddings that preserve invariances such as rotational symmetry, often projecting high-dimensional inputs into lower-dimensional spaces for downstream tasks [14, 18]. Dimensionality reduction enables tractable optimization but introduces geometric compression, where structurally meaningful separations risk collapse into statistically convenient proximities.

Yet, the reliance on Euclidean metrics in loss functions and similarity computations introduces biases, particularly in disordered or amorphous materials where local geometries deviate from flat spaces [8, 31]. In such systems, bonding irregularities and coordination variability produce curved relational neighborhoods that resist linear encoding.

Deep learning variants, including generative adversarial networks, have been applied to sample composition spaces inversely, but their Euclidean priors limit the fidelity of generated representations [13, 23]. Generated candidates may satisfy compositional constraints while violating latent structural feasibility embedded within curved manifolds.

Literature highlights the need for curvature-aware embeddings to better handle hierarchical features, such as in perovskites or high-entropy alloys, where traditional metrics undervalue long-range interactions [21, 25]. Embedding frameworks that incorporate hyperbolic or mixed-curvature geometries demonstrate improved capacity for preserving compositional taxonomies and phase hierarchies.

## AI-Guided discovery systems

AI integration has revolutionized discovery systems by enabling autonomous navigation of material landscapes through active learning and Bayesian optimization [5, 15]. These systems orchestrate predictive modeling, candidate prioritization, and experimental validation within closed computational loops.

Closed-loop experimentation couples predictive models with synthesis platforms, iteratively refining candidates based on feedback [16]. Such architectures compress discovery cycles from decades to iterative computational–experimental exchanges conducted over weeks or months.

High-throughput computations accelerate this by screening virtual libraries, but Euclidean distance metrics can misguide selection processes, favoring clusters that appear proximal yet are topologically distant [11, 12]. This misalignment is particularly pronounced near phase boundaries, where linear similarity gradients diverge from thermodynamic realities.

Uncertainty quantification plays a critical role here, with techniques like Gaussian processes estimating prediction variances; however, these assume linear separability incompatible with manifold structures [14, 28]. Consequently, epistemic uncertainty may be underreported in regions where curvature-induced distortion is highest.

Systems-level insights reveal trade-offs in computational steering, where over-reliance on flat metrics amplifies epistemic risks in extrapolation, potentially overlooking novel phases or compositions [17, 30]. Discovery systems may therefore converge prematurely within locally dense yet globally suboptimal design basins.

## Computational design paradigms

Inverse design paradigms shift from forward prediction to generating materials that meet specified criteria, leveraging machine learning to explore unconstrained spaces [2, 16]. This transition reframes AI from predictive assistant to generative architect.

Stochastic methods, such as genetic algorithms combined with neural networks, optimize designs, but Euclidean embeddings constrain the search to linear trajectories, missing curved pathways that could yield breakthroughs [13, 27]. As a result, optimization landscapes may contain inaccessible basins reachable only through curvature-respecting traversal.

In foundation models for science, pre-trained on diverse datasets, the paradigm emphasizes transfer learning across domains, yet the critique lies in how distance metrics influence fine-tuning, often propagating biases from Euclidean assumptions [8, 17]. Knowledge transfer may therefore replicate geometric distortions embedded in source datasets.

Multimodal integration further complicates this, as aligning simulation-derived features with experimental data requires metrics sensitive to underlying geometries [22, 29]. Without curvature harmonization, inverse design outputs risk feasibility gaps between predicted and synthesizable materials.

## Uncertainty & interpretability

Addressing uncertainty is paramount in materials AI, where models must convey confidence in predictions to inform decision-making [14, 19]. Predictive outputs increasingly function as experimental steering signals, making epistemic transparency essential.

Techniques incorporating physics-informed constraints enhance reliability, but Euclidean frameworks limit interpretability by obscuring manifold curvatures that underlie variance [11, 18]. When curvature distortion is hidden, uncertainty appears uniformly distributed rather than structurally localized.

Literature synthesizes that interpretability hinges on dissecting representation–inference interactions, revealing how distance distortions affect epistemic structures [6, 20]. Visualization of latent manifolds demonstrates that predictive confidence often correlates with geometric density rather than physical certainty.

In graph-based systems, this manifests as challenges in propagating uncertainties through layers, where non-

Euclidean considerations could offer clearer insights into model behaviors [26]. Curvature-aware propagation would allow uncertainty to diffuse along structurally meaningful pathways instead of uniformly across embeddings.

Overall, the synthesis underscores infrastructure trade-offs, advocating for interpretive lenses that prioritize geometric fidelity over simplistic metrics [4, 32]. The literature converges on the need to transition from flat similarity logics toward topology-preserving discovery paradigms capable of sustaining epistemic robustness in expanding materials design spaces.

## Proposed conceptual framework

To address the critiqued limitations of Euclidean distance metrics in data-driven materials discovery, we propose the Manifold-Aware Discovery Architecture (MADA), an original interpretive framework that reconceptualizes material spaces as curved manifolds rather than flat planes. MADA structures the discovery process into interconnected layers: data ingestion, representation embedding, inference navigation, and feedback integration, each informed by geodesic logics to preserve topological integrity. Rather than treating materials similarity as linear proximity, the framework interprets discovery as navigation across curvature-modulated knowledge terrains shaped by bonding energetics, crystallographic discontinuities, and multiscale structure–property couplings.

## Data ingestion layer — Curvature-aware substrate formation

At the data ingestion layer, MADA interprets multimodal inputs—such as compositional formulas, crystal structures, and property spectra—as points on a Riemannian manifold, where distances reflect intrinsic curvatures arising from atomic interactions and phase boundaries. This contrasts with Euclidean projections by employing sectional curvature measures to weigh feature importance, ensuring that sparse regions, like those in high-entropy materials, are not artificially compressed [7, 21].

Expanding this logic, ingestion is reframed as a geometric encoding act rather than a preprocessing step. Feature vectors are mapped through curvature-sensitive kernels that preserve thermodynamic discontinuities, allowing phase transitions, metastability basins, and polymorphic bifurcations to retain spatial separability. In this configuration, data imbalance is interpreted as curvature asymmetry rather than sampling deficiency, enabling structurally rare materials families to maintain epistemic volume within the manifold substrate. Consequently, ingestion becomes the first epistemic safeguard against representational flattening.

## Representation embedding layer — Non-euclidean learning architectures

The representation embedding layer builds upon this by utilizing hyperbolic or spherical geometries in neural architectures, allowing for hierarchical expansions that better accommodate exponential growth in chemical diversity [8, 31]. Here, graph neural networks are steered via curvature-adaptive aggregations, where message passing follows geodesic paths rather than straight-line norms, enhancing the capture of long-range dependencies without distortion [18, 26].

This embedding regime introduces curvature-conditioned representation scaling, where negatively curved latent regions expand compositional hierarchies, while positively curved pockets stabilize structurally dense clusters. Such geometric partitioning enables simultaneous encoding of periodic crystal families and non-periodic amorphous systems within a unified manifold. Moreover, curvature-regularized loss functions penalize distortive embeddings, ensuring that latent proximities remain physically interpretable rather than statistically convenient.

## Inference navigation layer — Geodesic steering logics

The inference navigation layer introduces discovery steering logics that prioritize manifold geodesics for similarity searches and optimization trajectories. In inverse design workflows, this means generating candidates along curved contours that align with physical constraints, mitigating biases in Euclidean-based sampling [13, 16].

Here, optimization is reconceptualized as trajectory shaping rather than point estimation. Gradient descent pathways are reparameterized along curvature fields, ensuring that search movements respect thermodynamic feasibility corridors. This prevents algorithmic shortcuts that traverse nonphysical regions of materials space.

Uncertainty quantification is reframed through Ricci curvature flows, providing epistemic risk structures that highlight regions of high distortion, thus guiding active

learning toward informative manifolds rather than misleading clusters [14, 19]. In this sense, uncertainty becomes geometrically locatable—manifesting not only as probabilistic spread but as structural deformation within the discovery topology.

## Feedback integration layer — Curvature realignment dynamics

Feedback loops close the architecture by coupling simulation–experiment dynamics, where discrepancies are interpreted as curvature mismatches, iteratively refining the manifold model to improve alignment across modalities [15, 22].

Within this layer, experimental validation functions as a topological correction mechanism. When predicted property gradients fail to align with empirical measurements, the manifold undergoes curvature recalibration, adjusting geodesic pathways and density distributions. This transforms feedback from simple model updating into structural manifold evolution.

Simulation pipelines, autonomous laboratories, and high-throughput screening platforms feed continuous corrections into the embedding geometry, enabling adaptive stabilization of discovery routes. Over successive cycles, the manifold transitions from speculative topology toward empirically anchored curvature, strengthening epistemic reliability.

## Computational workflow dynamics — Infrastructure and epistemic Trade-Offs

Computational workflow dynamics within MADA emphasize trade-offs in infrastructure scalability, such as the balance between embedding dimensionality and geodesic computation costs, fostering more robust AI pipelines [6, 12]. High-dimensional curvature embeddings enhance representational fidelity but introduce navigation overheads, necessitating optimized geodesic solvers and sparse manifold approximations.

Representation–inference interactions are highlighted through interpretive mappings that visualize how non-Euclidean metrics influence decision boundaries, offering systems-level insights into discovery efficiency [4, 20]. These mappings expose how curvature reshapes clustering thresholds, similarity basins, and optimization funnels, thereby revealing hidden steering logics within AI-guided materials exploration.

## Systems-level implication

Taken together, MADA reframes materials discovery as a topology-preserving epistemic system rather than a distance-minimization exercise. By embedding ingestion, learning, inference, and validation within curvature-aware geometries, the framework preserves intrinsic material relationships while mitigating distortions introduced by flat metric assumptions. This manifold-centric orientation enables more faithful navigation of vast chemical design spaces, aligning computational exploration with the true structural complexity of matter. As conceptualized in **Figure 1**, this framework unveils epistemic structures that traditional approaches overlook.

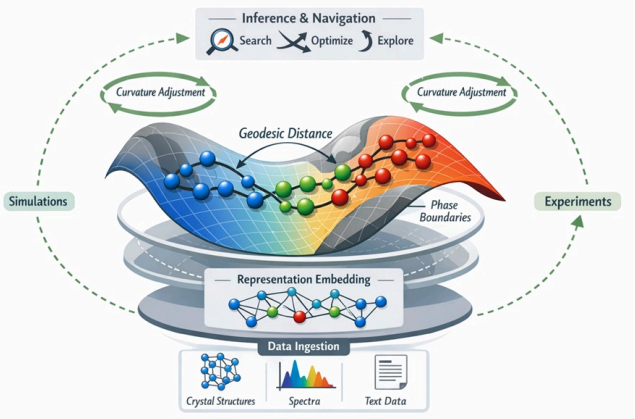


**Figure 1.** Schematic representation of the MADA framework as a layered manifold surface, showing geodesic arcs linking data points to discovery outcomes, iterative feedback arrows, and navigation paths that circumvent high-curvature zones.

## Analytical implications

The Manifold-Aware Discovery Architecture (MADA) offers interpretive lenses for dissecting the implications of non-Euclidean metrics in computational materials workflows, revealing how geometric fidelity influences data-to-discovery pipelines. By prioritizing geodesic distances over Euclidean norms, MADA highlights analytical trade-offs in representation learning, where curved embeddings can mitigate distortions in high-dimensional spaces, potentially refining the accuracy of property inferences without altering underlying datasets [7, 8]. This shift underscores systems-level insights into how manifold curvatures interact with

feature hierarchies, such as in crystalline versus amorphous materials, allowing for more nuanced clustering that respects topological boundaries [18, 31].

In terms of discovery steering logics, MADA interprets the navigation of material spaces as path-dependent processes, where Euclidean shortcuts may lead to epistemic pitfalls like false similarities in compositionally diverse alloys [21, 25]. Geodesic-aware steering, conversely, fosters computational dynamics that align better with physical realities, such as phase stability landscapes, enabling workflows to explore curved trajectories that uncover overlooked optima [13, 16]. Feedback loops within MADA amplify this by interpreting iterative refinements as curvature adjustments, which could enhance the coupling of high-throughput simulations with experimental validations, reducing discrepancies arising from flat metric assumptions [12, 15].

Epistemic risk structures emerge as a key implication, with MADA framing uncertainty not as isotropic variances but as directionally dependent on manifold geometry [14, 19]. This perspective reveals infrastructure trade-offs, such as the computational cost of embedding in non-Euclidean spaces versus the epistemic gains in interpretability, particularly in multimodal integrations where data modalities exhibit varying curvatures [20, 22]. For AI-guided systems, this implies refined uncertainty propagation in graph neural networks, where layer-wise geodesic aggregations could provide clearer insights into model vulnerabilities [6, 32]. Overall, these implications suggest a reevaluation of data infrastructures, advocating for tools that incorporate curvature metrics to support more resilient discovery ecosystems [9, 10].

# Results and Discussion

While the Euclidean critique embedded in MADA illuminates core limitations in current data-driven paradigms, it also invites broader discourse on the interplay between computational abstractions and materials reality. Traditional metrics, though computationally efficient, often impose artificial linearity on inherently curved spaces, as seen in the challenges of representing disordered systems or multi-scale phenomena [8, 31]. MADA's manifold lens encourages interpretive integrations that bridge these gaps, potentially informing the evolution of foundation models by emphasizing geometric priors in pre-training [8, 17].

However, implementing such frameworks conceptually raises questions about scalability in high-throughput environments, where geodesic computations might introduce overheads that trade off against rapid screening [11, 12]. Representation–inference interactions under MADA suggest that while curvature awareness enhances fidelity, it necessitates adaptive architectures capable of handling variable geometries, drawing from advances in generative models for inverse design [13, 23]. Uncertainty and interpretability remain pivotal, with MADA's risk structures highlighting the need for metrics that quantify epistemic distortions, akin to how physics-informed constraints temper model outputs [11, 18].

In the context of autonomous systems, the framework's feedback dynamics underscore the importance of closed-loop adaptability, where manifold refinements could optimize experiment-simulation couplings amid noisy data [15, 16]. Yet, epistemic trade-offs persist, such as balancing global topology capture with local feature precision, which could influence the design of multimodal datasets [22, 29]. Ultimately, MADA's insights advocate for a paradigm where computational steering is geometry-informed, fostering innovations in materials informatics without empirical mandates [5, 6].

# Conclusion

The critique of Euclidean distance metrics through the Manifold-Aware Discovery Architecture underscores a pivotal conceptual pivot in data-driven materials discovery, emphasizing the need for geometric fidelity to navigate complex material spaces effectively. By interpreting pipelines through manifold lenses, MADA reveals analytical and systemic enhancements that could refine representation learning, steering logics, and epistemic structures, ultimately advancing computational infrastructures in materials engineering. This framework positions future workflows toward more interpretive and integrative approaches, promising accelerated discoveries in critical domains while highlighting inherent trade-offs in AI applications.

# Acknowledgements

None

# Conflict of interest

None

## Financial support

None

## Ethics statement

None

Received: 12 Jun 2021 Revised: 12 Aug 2021 Accepted: 29 Nov 2021
Published online: 18 March 2022



## References

Ramprasad R, Brgoch J, Lordi V. Machine learning in materials informatics: Recent applications and prospects. npj Comput Mater. 2017;3(1):54. https://doi.org/10.1038/s41524-017-0056-5.

Liu Y, Zhao T, Ju W, Shi S. Materials discovery and design using machine learning. J Materiomics. 2017;3(3):159-77. https://doi.org/10.1016/j.jmat.2017.08.002.

Takahashi K, Tanaka Y. Material synthesis and design from first principle calculation and machine learning. Comput Mater Sci. 2016;112:364-77. https://doi.org/10.1016/j.commatsci.2015.10.050.

Zhou T, Song Z, Sundmacher K. Big data creates new opportunities for materials research: A review on methods and applications of machine learning for materials design. Engineering. 2019;5(6):1017-26. https://doi.org/10.1016/j.eng.2019.02.011.

Zhang Y, Ling C. A strategy to apply machine learning to small datasets in materials science. npj Comput Mater. 2018;4(1):25. https://doi.org/10.1038/s41524-018-0081-6.

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. https://doi.org/10.1038/s41524-019-0221-0.

Goodall REA, Lee AA. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. Nat Commun. 2020;11(1):6280. https://doi.org/10.1038/s41467-020-19964-7.

Chen C, Zuo Y, Ye W, Li X, Ong SP. Learning properties of ordered and disordered materials from multi-fidelity data. Nat Comput Sci. 2021;1(1):46-53. https://doi.org/10.1038/s43588-020-00002-x.

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(1):138. https://doi.org/10.1038/s41524-020-00406-3.

Ward L, Dunn A, Faghaninia A, Zimmermann NER, Bajaj S, Wang Q, et al. Matminer: An open source toolkit for materials Pilania G, McClellan KJ, Stanek CR, Uberuaga BP. Physics-informed machine learning for inorganic scintillator discovery. J Chem Inf Model. 2018;58(6):1345-55. https://doi.org/10.1021/acs.jcim.8b00131.

Ward L, Dunn AR, Faghaninia A, Zimmermann NER, Bajaj S, Wang Q, et al. Matminer: An open source toolkit for materials data mining. Comput Mater Sci. 2018;152:60-9. https://doi.org/10.1016/j.commatsci.2018.05.018.

Wang AY-T, Murdock RJ, Kauwe SK, Bukarjoo AT, Knol A, Müller K, et al. LaQA: A latent representation model for materials prediction using quantum-inspired algorithm. Comput Mater Sci. 2021;188:110228. https://doi.org/10.1016/j.commatsci.2020.110228.

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. https://doi.org/10.1038/s41524-020-00352-0.

Chen L, Tran H, Batra R, Kim C, Ramprasad R. Machine learning models for the prediction of energy, forces, and stresses for molecules and materials. npj Comput Mater. 2021;7(1):19. https://doi.org/10.1038/s41524-021-00489-6.

Omee SS, Louis S-Y, Hu J. Automatized assessment of protective group reactivity for synthetic chemistry. npj Comput Mater. 2021;7(1):114. https://doi.org/10.1038/s41524-021-00585-7.

Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: Generative models for matter engineering. Science. 2018;361(6400):360-5. https://doi.org/10.1126/science.aat2663.

Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. Nature. 2023;624(7990):80-5. https://doi.org/10.1038/s41586-023-06735-9.

Glielmo A, Zeni C, De Vita A. Efficient nonparametric n-body force fields from machine learning. Phys Rev B. 2018;97(18):184307. https://doi.org/10.1103/PhysRevB.97.184307.

Bartel CJ, Sutton C, Goldsmith BR, Oganov AR, Yu Y, Zhu L, et al. A critical examination of compound stability predictions from machine-learned formation energies. npj Comput Mater. 2020;6(1):97. https://doi.org/10.1038/s41524-020-00362-y.

Antoniuk ER, Suh PJ, Chen Q, Goodall REA, Browning AR, Jain A, et al. A general-purpose material property data extraction system from literature powered by natural language processing. Comput Mater Sci. 2022;215:111806. https://doi.org/10.1016/j.commatsci.2022.111806.

Kaufmann K, Maryanovsky D, Mellor WM, Zhu C, Rosengarten AS, Vecchio KS. Discovery of high-entropy ceramics via machine learning. npj Comput Mater. 2020;6(1):42. https://doi.org/10.1038/s41524-020-0317-6.

Kim E, Huang K, Saunders S, McCallum A, Ceder G, Olivetti E. Materials synthesis insights from scientific literature via text extraction and machine learning. Chem Mater. 2017;29(21):9436-44. https://doi.org/10.1021/acs.chemmater.7b03570.

Dan Y, Zhao Y, Li X, Li S, Hu M, Hu J. Low-data deep learning for the inverse design of organic photovoltaic acceptors. npj Comput Mater. 2021;7(1):23. https://doi.org/10.1038/s41524-021-00493-w.

Isayev O, Oses C, Toher C, Gossett E, Curtarolo S, Tropsha A. Universal fragment descriptors for predicting properties of inorganic crystals. Nat Commun. 2017;8(1):15679. https://doi.org/10.1038/ncomms15679.

Ward L, Wolverton C. Atomistic calculations and materials informatics: A review. Curr Opin Solid State Mater Sci. 2017;21(3):167-76. https://doi.org/10.1016/j.cossms.2016.07.002.

Zeni C, Bryden A, Pini G, Bachmann R, Polyudov G, Pironti A, et al. Graph neural networks for simulating crack coalescence and propagation in brittle materials. Comput Methods Applied Mech Eng. 2022;395:115051. https://doi.org/10.1016/j.cma.2022.115051.

Meredig B, Agrawal A, Kirklin S, Saal JE, Doak JW, Thompson A, et al. Combinatorial screening for new materials in unconstrained composition space with machine learning. Phys Rev B. 2014;89(9):094104. https://doi.org/10.1103/PhysRevB.89.094104.

Hansen K, Biegler F, Ramakrishnan R, Pronobis W, Von Lilienfeld OA, Müller KR, et al. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. J Phys Chem Lett. 2015;6(12):2326-31. https://doi.org/10.1021/acs.jpclett.5b00831.

Liu R, Tawa G, Wallqvist A. Locally linear embedding (LLE) for MRI based Alzheimer's disease classification. Neuroimage. 2013;83:148-57. https://doi.org/10.1016/j.neuroimage.2013.06.033.

Faber FA, Lindmaa A, Von Lilienfeld OA, Armiento R. Machine learning energies of 2 million elpasolite (ABC2D6) crystals. Phys Rev Lett. 2016;117(13):135502. https://doi.org/10.1103/PhysRevLett.117.135502.

Pilania G, Mannodi-Kanakkithodi A, Uberuaga BP, Ramprasad R, Gubernatis JE, Lookman T. Machine learning bandgaps of double perovskites. Sci Rep. 2016;6:19375. https://doi.org/10.1038/srep19375.

Ghiringhelli LM, Vybiral J, Levchenko SV, Draxl C, Scheffler M. Big data of materials science: critical role of the descriptor. Phys Rev Lett. 2015;114(10):105503. https://doi.org/10.1103/PhysRevLett.114.105503.