



Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/244341/>

Version: Published Version

Article:

Liu, X., Anjah, C., Jolly, B.E.E. et al. (2026) Generative and multimodal AI for materials prediction and design: progress, challenges, and perspectives. Journal of Physics: Materials. ISSN: 2515-7639

<https://doi.org/10.1088/2515-7639/ae93fd>

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

ACCEPTED MANUSCRIPT • OPEN ACCESS

Generative and multimodal AI for materials prediction and design: Progress, challenges, and perspectives

To cite this article before publication: Xianyuan Liu *et al* 2026 *J. Phys. Mater.* in press <https://doi.org/10.1088/2515-7639/ae93fd>

Manuscript version: Accepted Manuscript

Accepted Manuscript is “the version of the article accepted for publication including all changes made as a result of the peer review process, and which may also include the addition to the article by IOP Publishing of a header, an article ID, a cover sheet and/or an ‘Accepted Manuscript’ watermark, but excluding any other editing, typesetting or other changes made by IOP Publishing and/or its licensors”

This Accepted Manuscript is © 2026 The Author(s). Published by IOP Publishing Ltd.



As the Version of Record of this article is going to be / has been published on a gold open access basis under a CC BY 4.0 licence, this Accepted Manuscript is available for reuse under a CC BY 4.0 licence immediately.

Everyone is permitted to use all or part of the original content in this article, provided that they adhere to all the terms of the licence <https://creativecommons.org/licenses/by/4.0>

Although reasonable endeavours have been taken to obtain all necessary permissions from third parties to include their copyrighted content within this article, their full citation and copyright line may not be present in this Accepted Manuscript version. Before using any content from this article, please refer to the Version of Record on IOPscience once published for full citation and copyright details, as permissions may be required. All third party content is fully copyright protected and is not published on a gold open access basis under a CC BY licence, unless that is specifically stated in the figure caption in the Version of Record.

View the [article online](#) for updates and enhancements.

Generative and multimodal AI for materials prediction and design: Progress, challenges, and perspectives

Xianyuan Liu^{1,2}, Charles Anjah^{1,2}, Benjamin E. Jolly^{3,4}, Jonathon F. S. Markanday⁵, Joshua Berry³, Haolin Wang^{1,2}, Nicola A. Morley³, Robert D. J. Oliver³, Alexandra J. Ramadan⁶, Delvin Ce Zhang¹, Katerina A. Christofidou^{3,4*}, Haiping Lu^{1,2,*}

¹School of Computer Science, University of Sheffield, Sheffield, UK

²Centre for Machine Intelligence, University of Sheffield, Sheffield, UK

³School of Chemical, Materials and Biological Engineering, University of Sheffield, Sheffield, UK

⁴Henry Royce Institute, Royce Discovery Centre, University of Sheffield, Sheffield, UK

⁵Materials Nexus Ltd., Salisbury House, Cambridge, UK

⁶School of Mathematical and Physical Sciences, University of Sheffield, Sheffield, UK

* Author to whom any correspondence should be addressed.

E-mail: k.christofidou@sheffield.ac.uk; h.lu@sheffield.ac.uk

Keywords: Multimodal Learning, Generative AI, Materials Design

Abstract

Artificial intelligence (AI) is accelerating materials prediction and design by enabling efficient exploration of chemical and structural spaces, with particular promise for novel materials discovery. However, novelty in materials discovery encompasses chemical plausibility, structural distinctiveness, property relevance and experimental realisability, making AI-driven novelty claims difficult to substantiate. We introduce a materials property hierarchy, from intrinsic, composition-determined properties to extrinsic, processing-dependent performance, to clarify deployment constraints and distinguish structural, physical and deployment novelty. This framework motivates an evidence-based view of multimodal materials data spanning chemical composition, microstructure, processing, and testing and characterisation, showing that current evidence remains concentrated in composition and idealised structure while heterogeneous, under-represented and weakly integrated modalities limit support for physical and deployment novelty. It also highlights the limitations of benchmarks based mainly on computational labels and proxy novelty criteria. Community-wide standards for data collection, modality alignment and evidence synthesis are needed to support multimodal data construction, process-aware multimodal modelling, feasibility-first generative modelling and deployment-aware benchmarking, so that generative and multimodal AI can design experimentally realisable materials with defensible scientific and practical novelty.

1 Introduction

Artificial intelligence (AI) has emerged as a practical engine for materials research, enabling the design and optimisation of candidate systems by navigating vast, complex compositional spaces to support exploration beyond manual reasoning or trial-and-error [1, 2]. Figure 1 shows that generative and multimodal AI have recently expanded rapidly in the materials research domain, both in property prediction and materials design.

Defining novelty in materials discovery remains challenging. Novelty is established not solely from unfamiliar composition or structure, but also from chemical plausibility, physical behaviour, and experimental realisability. The criteria for establishing novelty depend on the design objective, and the evidence required to support novelty is distributed across materials data. Materials data are inherently multimodal, spanning chemical composition, crystalline phase, process history, characterisation data, and the technical narratives contained in laboratory records and the scientific literature. These diverse modalities provide complementary insights, highlighting the relationships between processing, structure, property, and performance that are fundamental to materials science [3].

Historically, materials discovery has largely relied on structured data representations such as chemical formulae [4, 5, 6], crystal structures [7, 8, 9, 10], and graphs that represent local atomic environments and lattice connectivity [11, 12, 13]. These representations have been highly effective for training models that map structural features to target properties [14, 15, 16]. However, these advances do not resolve the question of novelty, because material behaviours in real-world applications often depend on processing accessibility, microstructural realisation, and deployment constraints that are only partially captured by idealised structural representations.

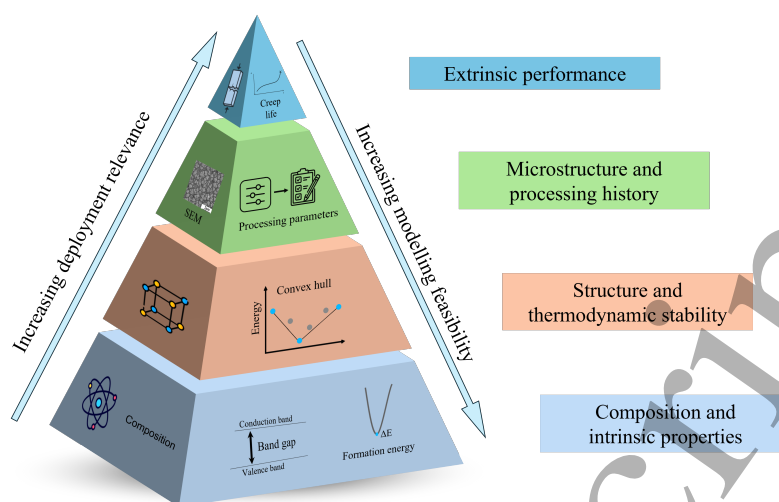


Figure 2: Hierarchical pyramid of materials property, from composition-determined intrinsic properties to processing-dependent extrinsic performance.

Intrinsic properties are primarily determined by the ground-state chemistry and atomic arrangement. When composition and phase are fixed, these properties are often relatively insensitive to processing history or microstructural variation. This is why crystal structure models and density functional theory (DFT) have been effective for predicting many electronic and thermodynamic quantities [27, 28, 29, 30].

Extrinsic properties, by contrast, determine the engineering performance under practical conditions. Fracture toughness, fatigue life, and many functional responses depend not only on composition and crystal structure, but also on defects, interfaces, grain structure, phase morphology, and processing history. These properties are path-dependent. Two materials with identical compositions and crystal structures can exhibit different mechanical or functional behaviours if they are processed differently. This path dependence highlights the limitations of unimodal or few-modality models that rely primarily on crystal structure, and it motivates the use of multimodal data that capture the full processing-structure-property chain.

This hierarchy is not proposed for a rigid boundary interpretation. In structural metallic alloys, the distinction is relatively clear. Intrinsic properties provide the thermodynamic foundation, while extrinsic performance is largely governed by microstructure and processing. However, in functional materials, intrinsic and extrinsic properties are often more strongly coupled. For example, in thermoelectrics, piezoelectrics, and battery electrode materials, key properties depend on both the electronic structure and microstructural features. Nevertheless, the hierarchy remains useful because it clarifies the organisation of material properties and behaviours and their relationship to materials data.

2.2 Deployment constraints

When the design objective shifts from intrinsic properties to extrinsic performance, deployment constraints emerge, including thermodynamic stability, processing accessibility, microstructural realisation, and practical synthesis constraints.

Computational materials discovery has traditionally focused on identifying thermodynamically stable phases. However, many technologically important materials rely on metastable structures that emerge through controlled processing routes. Their value arises not from occupying the global energy minimum, but from being accessible via specific synthesis routes and retaining useful properties in practice. Distance from the convex hull is a stability-related measure for synthetic feasibility commonly used in generative models for crystal structure. It provides a useful first check, as candidates far above the hull are unstable and therefore less likely to be synthesised. However, it remains a rough, indirect proxy because stability is only one of the factors influencing synthesizability. Experimentally important materials often lie in metastable regions of the energy landscape, whereas some predicted ground states remain difficult to realise [31].

The importance of metastability is evident across multiple material domains. In structural steels, useful combinations of high strength and ductility arise from microstructures produced through controlled thermomechanical processing rather than from equilibrium phase selection alone. Similarly, amorphous silicon and certain thin-film semiconductors often derive their technological value from non-equilibrium states stabilised by vapour deposition or rapid quenching [32, 33, 34]. In such cases, the relevant design target is not only a stable crystal structure, but also a process-accessible state with useful behaviour.

More broadly, material behaviour is also influenced by coupled processing-structure-property relationships. Microstructural features, such as phase distribution, grain morphology, and defect structures,

strongly influence performance. Practical deployment also imposes synthesis constraints, including elemental availability, regulatory requirements, and compositional tolerances [2, 35, 36, 37]. These factors often determine whether a computational proposal remains a theoretical possibility or becomes a real material.

In this Perspective, deployment is considered primarily in terms of synthesisability and experimental realisability, where synthesisability refers to whether a material can be produced through a plausible processing route, and experimental realisability refers to whether the resulting sample reproduces the targeted composition, structure, and microstructure under measurable conditions.

2.3 Materials novelty taxonomy

The materials property hierarchy and deployment constraints shape the meaning of novelty in materials discovery. Statistical distance in composition or latent space is a reasonable screening criterion for intrinsic properties, but it does not measure extrinsic performance. A proposed material may sit far from any known composition and still be unstable, inaccessible, or functionally unremarkable, while a near-identical composition may behave very differently if it is processed differently. Therefore, novelty is not a single criterion. Following the materials property hierarchy introduced in Sec. 2.1, we organise material novelty in a hierarchical framework.

The first level is structural novelty, where a material occupies a previously unexplored region of compositional, crystallographic, or phase space. The second is physical novelty, where it exhibits properties, phase behaviour, or microstructural characteristics not observed in existing systems under comparable conditions. The third is deployment novelty, where it delivers performance that existing materials cannot achieve under realistic synthesis, manufacturing, or operating constraints.

These levels are relevant to different design objectives. Structural novelty may be sufficient for exploring new chemical or structural space, while physical novelty is important for identifying new intrinsic behaviour or distinct phase responses. Deployment novelty is required for designing experimentally realisable and practically useful materials.

2.4 Materials novelty evaluation

The evaluation at each novelty level requires different forms of material evidence, and also differs in how novelty can be established in practice. Structural novelty is the most tractable and can be evaluated against composition and structure alone, for example, using the distance to the nearest known material in a defined descriptor or structural space [38, 39], combined with a basic stability check such as distance from the convex hull [31, 40]. The definition of “known” depends on the criterion used for comparison.

Structure-matching tools, such as pymatgen’s StructureMatcher [41], are commonly used, while recent Pointwise Distance Distributions [42] offer an alternative, identifying near-duplicates by comparing crystals as unordered point sets under isometry. Such comparisons are also sensitive to how disorder and reference structures are represented, as reflected in the recent discussion about the novelty of the MatterGen-proposed Ta-Cr-O candidate [43].

Physical novelty evaluation requires a criterion in property rather than in input space. Based on a trusted predictor or measurement, a candidate material is physically novel only if its behaviour falls outside the reference distribution of known materials under comparable conditions, not only outside the training set [25, 26]. Predicted composition and crystal structure may be distinct and yet not synthesisable, in which case the associated behaviour is never realised. A property predicted from an idealised structure remains only a candidate for physical novelty until that behaviour is demonstrated in a material that can be made in practice [31]. In structural steel, for example, behaviour is determined jointly by the phases present, the microstructure produced by thermomechanical processing, and the resulting properties under relevant conditions, such as the combination of strength and ductility. A new steel composition occupying an unexplored region of the alloying space is only structurally novel. The same steel is physically novel only if a realisable, processed form exhibits a strength-ductility combination outside the range achieved by existing steels under comparable conditions, rather than a combination inferred from an idealised structure alone.

Deployment novelty sets a higher bar because its decisive properties are often extrinsic, path-dependent, and time-dependent. Evaluating these properties may require longer and more complex tests, making them difficult to incorporate into iterative model development. Surrogate proxies and early-stage indicators can support evaluation, but deployment novelty ultimately requires a demonstrable advantage over existing materials under realistic synthesis and operating conditions. For structural applications in particular, this advantage must hold not only in the as-synthesised or as-processed state but also throughout the service life, since properties can degrade through in-service exposure, including oxidation, fatigue, and corrosion. Therefore, deployment novelty extends to retained functionality under service-relevant ageing, rather than initial performance alone. This emphasis on retained functionality also connects to broader discussions of the distinction between a compound and a material, in which functionality and potential utility are central [44]. Establishing physical and deployment novelty requires information beyond composition and

Table 1: Multimodal representations of materials data: A metallic alloy example

Data category	Data records	AI-ready encodings
Chemical composition	Concentrations of each elemental constituent; Formula and stoichiometry; Precursor purity, ratios and composition ranges	Element-fraction vectors; Descriptor vectors; Element tokens with learned embeddings; Serialised composition text
Microstructure	Microscopy images; Chemical information and spectroscopy data; Orientation and phase maps; Segmentations and region annotations	Patch or image embeddings; Hyperspectral images and embeddings; Multi-scale texture and morphology features; Region-level attributes; Adjacency graphs between regions and/or phases
Processing	Manufacturing process parameters; Temperature, pressure, and time logs; In situ sensor streams	Step sequences or action graphs; Parameter vectors with protocol tags; Time-series embeddings with event markers
Testing and characterisation	Measured properties and curves; Numerical mechanical property data; Imaging from broken specimens; Instrument settings and analysis metadata	Curve and spectrum embeddings; Peak and shape descriptors; Metadata-aware features; Fidelity tags for simulation vs experiment

structure, including processing history, microstructure, synthesis records and measured performance. Since this information is distributed across different data types, multimodal data and AI workflows become essential to developing AI models for novel materials design.

3 Multimodal materials data

The hierarchy in Sec. 2 shows that no single data type can fully characterise a material. Composition defines the chemistry, but processing, microstructure and characterisation together determine how that chemistry translates into observable behaviour. This section organises materials data into these four categories and introduces how they are currently represented, extracted and distributed.

3.1 Data modalities and representations

Materials data are inherently multimodal and can be broadly organised into four categories based on their data sources: **chemical composition, microstructure, processing, and testing and characterisation** [45]. These categories are consistent with the broader materials property hierarchy. Each category may contain multiple modalities, and a single modality may contribute to several categories. Table 1 presents multimodal representations of metallic alloys.

Chemical composition captures the elemental identities and proportions of a material and therefore constrains phase stability and intrinsic behaviour [45]. It is commonly represented as chemical formulae, atomic-fraction vectors, or human-designed descriptors based on elemental properties [46, 47, 48, 49]. Composition is one of the most widely used material descriptors because it is specified before synthesis and remains relatively consistent across datasets. It is typically reported in a structured numerical form but can also be extracted from scientific texts and experimental reports [50, 51].

Microstructure captures geometric and crystallographic information beyond that captured by composition alone. At the atomic scale, structural information is commonly represented using crystal structures encoded in formats such as Crystallographic Information Files (CIF), as atomic graphs [11, 12], or as 3D point clouds [52]. At larger length scales, microstructure comprises phase distributions, grain sizes, and crystallographic alignment, which are measured using microscopy and diffraction techniques and represented as images, voxel grids, or spatial fields [53, 54, 55].

Processing information records the synthesis routes used to produce the material, which determine phase transformation kinetics and the emergence of non-equilibrium states. This information varies significantly across material domains. Processing information is represented as synthesis text, equipment-generated time series, or structured provenance records that capture dependencies between fabrication steps [51, 56, 57, 58, 59, 60].

Testing and characterisation observations record the material behaviour under external stimuli and operating conditions. These observations include instrument- or simulator-level outputs, such as mechanical, electrical, magnetic, or spectroscopic responses [61, 62]. As they connect material state to measured function, they are essential for validating whether a proposed material meets its intended performance requirements [63, 64].

Figure 3a shows the annual growth in publications that incorporate these four categories. Their representation remains uneven, as shown in Fig. 3b. Testing and characterisation dominate, while microstructure has expanded in recent years. Composition remains the smallest component throughout the period, despite appearing in most multimodal studies, which reflects that it is often used as baseline information combined with other data types rather than analysed as a primary data input.

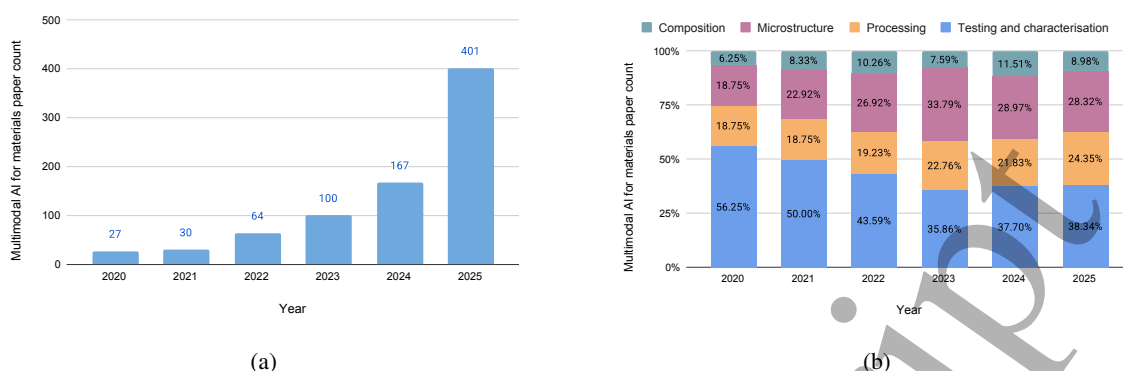


Figure 3: Multimodal AI for materials publications by data category from 2020 to 2025. (a) The annual number of publications that match the multimodal AI for materials query and involve at least one of four materials data categories. (b) Relative proportion of the data categories each year, where publications spanning multiple categories are counted in each. Publication data were retrieved from the Elsevier Scopus database on 24 June 2026. See Supplementary Information Section 1 for details of the trend analysis.

3.2 Multimodal data extraction

Across all four categories, materials data are rarely available in a structured, AI-ready form at the point of generation. Scientific literature remains one of the most important sources for data extraction, but relevant information is often dispersed across text, tables and figures, making acquisition and curation a major bottleneck. Literature-mining pipelines, such as ChemDataExtractor [65, 66], use named entity recognition and rule-based parsing to extract chemical entities and properties into structured databases. ChemDataExtractor remains widely used for its high precision and reproducibility, but its reliance on hand-crafted rules requires substantial domain expertise when extending it to new material classes. More recently, LLMs have helped to reduce this manual effort and improve extraction flexibility, including conversational prompting to extract and self-verify properties [67] and fine-tuned LLMs for structured entity and relation extraction [68]. Recent approaches are pushing extraction towards multimodal and larger-scale settings. LeMat-Synth [60] combines LLMs and vision-language models to extract synthesis procedures and performance data from both text and figures. MatSKRAFT [69] uses constraint-driven graph neural networks (GNNs) to assemble large materials knowledge bases from scientific tables. These approaches determine which modalities can be collected at scale and, consequently, shape the data foundation available for AI model development.

3.3 Coverage, imbalance, and heterogeneity

Despite these advances, multimodal materials data analysis faces several challenges, including limited coverage, modality imbalance, missing negative results, and heterogeneity across laboratories and measurement pipelines [70, 71]. These challenges constrain both the relationships that multimodal models can learn and their ability to generalise reliably beyond the training data.

Materials datasets are often scarce and concentrated in specific compositional or processing subspaces because synthesis and characterisation are expensive and time-consuming [6, 70, 72, 73, 74, 75, 76]. The imbalance exists within the synthesis and characterisation data. For example, microstructural imaging and rapid measurements, such as hardness, are widely reported, whereas detailed processing metadata, failed synthesis records, and long-duration performance measurements are much less available [72, 77, 78, 79]. In metallurgy and alloy development, long-term tests, such as creep, are particularly difficult to collect at scale because a single experiment may require 10^3 - 10^4 hours [80]. Consequently, available datasets usually capture only part of the processing-structure-property relationships, and are often dominated by the most accessible data types, which may not be the most informative for deployment-relevant behaviour.

This imbalance is also visible in widely used resources. Large open databases, such as the Materials Project [27], OQMD [28, 29] and JARVIS [30], provide large-scale, AI-ready data on compositions, idealised crystal structures and DFT-derived intrinsic properties, but offer limited coverage of processing and characterisation information [81]. Microstructures are expanding rapidly but remain fragmented across smaller repository depositions and literature-mined collections [82]. More comprehensive multimodal and processing-aware resources may exist on commercial materials information platforms, but they remain proprietary and largely inaccessible to open model development.

A related challenge is the systematic absence of negative or null results, namely “dark data”. Failed syntheses, unsuccessful processing routes, and compositions that do not produce the intended phase or property are rarely reported, even though they are important for constraining design space and defining feasibility boundaries [83, 84, 85].

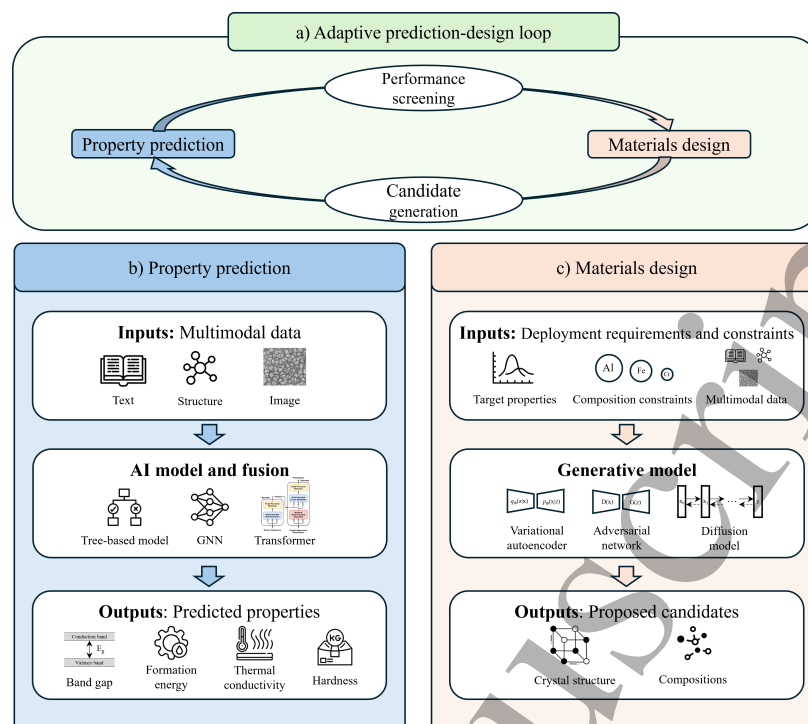


Figure 4: Generative and multimodal AI for materials discovery, including adaptive prediction-design loop (a), property prediction (b), and materials design (c).

Materials datasets are also highly heterogeneous across laboratories, instruments, and analysis pipelines. Experimental outputs depend strongly on sample preparation, instrument calibration, and downstream processing, while quality-control and reporting practices vary substantially between laboratories [86]. Without consistent metadata standards and detailed reporting of experimental conditions, it becomes difficult to compare results across sources or to distinguish physically meaningful variation from artefacts introduced during acquisition and analysis.

Taken together, these challenges show that multimodal materials datasets are not only limited in quantity but also uneven in coverage, selective in reporting, and variable in quality across sources.

4 Multimodal learning and generative AI for materials

The prior sections demonstrate that designing novel materials requires broader evidence from the materials property hierarchy, and that this evidence is distributed across heterogeneous data sources. Current AI for materials research can be organised around two coupled tasks: property prediction and materials design. As shown in Fig. 4, these tasks form an adaptive prediction-design loop in which candidate generation and performance screening iteratively inform each other. Within this loop, multimodal learning provides the foundation for extracting, aligning and fusing information across heterogeneous materials data, while generative AI can use these representations to propose new compositions, structures, or processing routes. Knowledge graphs and other knowledge-integration frameworks can effectively learn fragmented relational knowledge from the literature and experiments to support prediction and design. The following reviews recent progress in multimodal representation learning and fusion, knowledge integration, and generative AI for materials prediction and design.

4.1 Multimodal representation learning for materials prediction

Multimodal learning begins by encoding heterogeneous materials data into AI-ready representations, and different data types usually require different encoders.

Chemical composition can be represented either through handcrafted elemental descriptors used with conventional machine learning models [87] or through neural networks that learn directly from formula tokens and fractional amounts [88, 89]. Crystal structures are commonly encoded as periodic graphs for GNNs that aggregate local coordination environments into structure-aware representations [11, 12, 13]. Because composition and crystal structure are generally more accessible and more consistently curated than processing or characterisation data, these methods are widely used across many materials domains.

Processing information, such as synthesis protocols or thermomechanical histories, can be represented as ordered sequences or provenance graphs extracted from laboratory records and scientific

literature [59, 90, 91]. Microstructural observations are commonly encoded using vision models, including convolutional neural networks and vision transformers, to extract quantitative representations from microscopy images and orientation maps [53, 54, 55, 77]. One-dimensional experimental signals, such as X-ray diffraction patterns, can be processed with convolutional or attention-based sequence models [63, 64, 92, 93]. Textual records can be analysed using domain-specific language models to extract entities, relationships, and synthesis-relevant constraints [94, 95].

Recent work has also explored LLMs as a more unified interface for heterogeneous materials data. In these approaches, non-text inputs are converted into structured textual descriptions that can be processed within a shared language representation. This can simplify multimodal pipelines, although it may also sacrifice numerical precision or spatial detail [96].

Once modality-specific representations are learned, the next challenge is to align and fuse them into a shared representation that preserves complementary information across modalities. Cross-modal alignment aims to ensure that embeddings from different modalities correspond to the same material state, often using contrastive learning to pull matched pairs together and push mismatched pairs apart. In practice, this step is difficult, as modalities can differ in completeness, resolution, invariance, and processing dependency. Textual descriptions can be incomplete or ambiguous, and experimental measurements may correspond to different processing states. Fusion then combines aligned representations to make predictions. Simple concatenation remains a common baseline, while more advanced methods use attention or gating mechanisms to weight modalities based on their information content and reliability.

Multi-fidelity learning further extends this framework by integrating computational and experimental materials data, which often provide correlated but systematically biased views of material behaviour [97]. This is especially relevant in materials research, where broad simulation coverage coexists with smaller but higher-fidelity experimental datasets.

Overall, multimodal representation learning provides the predictive foundation for linking composition, structure, processing, and observation within a common modelling framework. By integrating complementary views of material state and behaviour, these methods can yield more robust predictions than unimodal models alone, and the multimodal representations subsequently serve as the conditioning signal that enables generative models to propose candidates compatible with these contexts (Sec. 4.3).

4.2 Knowledge integration via materials knowledge graphs

While multimodal representation learning focuses on encoding materials data into shared representations, scientific knowledge is often dispersed across literature, databases, and experimental records, rather than appearing as aligned training pairs. This has motivated growing interest in materials knowledge graphs, which provide a structured way to integrate materials data with broader scientific context [69, 98, 99, 100, 101, 102]. Materials knowledge graphs organise connected entities such as materials, properties, processes, synthesis and characterisation methods, phase labels, applications, and scientific documents within a common relational framework [98]. This graph structure provides a way to represent a scientific context that is difficult to capture from composition or structure alone.

Materials knowledge graphs can contribute at three levels. First, their relational structure captures connections between materials, properties, processes, and other scientific entities that are not available from internal composition or crystal structure alone, thereby supporting higher-order reasoning and more grounded property prediction [103]. Second, knowledge graphs can serve as an external knowledge source for materials language models, allowing updated information to be retrieved at inference time rather than relying entirely on static training data [19, 104, 105, 106, 107, 108, 109, 110]. Third, by linking structured data, text, and experimental records within a shared framework, they provide a practical route for integrating fragmented multimodal scientific knowledge [111, 112].

Overall, materials knowledge graphs support relational reasoning, knowledge grounding, and heterogeneous data integration, connecting composition and structure to the synthesis and characterisation context that often determines whether a material is scientifically meaningful and practically relevant.

4.3 Generative AI for multimodal materials design

Multimodal representations and knowledge structures provide the basis for generative modelling. The encoders and fusion strategies introduced in Sec. 4.1 for property prediction return here in a different role: they define the conditioning space used to propose new candidates. This allows generated outputs to be constrained not only by chemical validity but also by the processing, microstructural, and performance contexts associated with the design target. The following reviews recent generative architectures that use multimodal conditioning to move from prediction to candidate generation, and discusses the constraints used in the generation process.

Generative AI is reshaping materials discovery by proposing new compositions, structures, and processing-relevant candidates rather than only predicting the properties of known materials [113].

Generative models support both unconditional and conditional generations. Unconditional models explore the space of chemically valid materials without predefined targets. Conditional models guide generation towards desired properties or synthesis constraints, such as band gap, crystal symmetry, or synthesis-related requirements.

Several generative families have been explored for these tasks [2, 114, 115]. Variational autoencoders learn continuous latent spaces that support interpolation and optimisation. Generative adversarial networks enable fast sampling but often suffer from instability. Diffusion models iteratively refine noisy inputs into structured outputs and have become prominent for crystal generation. LLMs provide a complementary interface by treating materials as tokenised sequences, such as formulae, crystal descriptions, or synthesis protocols, allowing generation to be conditioned on textual or structured prompts [116]. LLMs for materials generation range from generating chemical formulae [105] to crystal structures as text [40]. However, their use also raises concerns about data leakage from large-scale pretraining and hallucinations that can introduce invalid structures, thereby motivating fine-tuning, validity checks, and post-generation filtering. These methods aim to move from passive prediction to active proposal of candidate materials.

Recent progress has been driven particularly by conditional generation. In materials design, useful conditions often combine hard constraints, such as stoichiometry, symmetry, or lattice consistency, with soft objectives, such as stability proxies, target property ranges, or dopability heuristics. Diffusion-based frameworks have proven flexible in this setting because constraints can be incorporated through training, guided sampling, or constrained decoding [2, 10, 117]. Related methods also treat generation as a constrained completion problem, for example, by reconstructing missing structural information from partial crystallographic or experimental evidence [118]. A further route is external-objective guidance, in which a pretrained generator is steered at sampling time using a surrogate model or reward, without retraining the conditional model. This idea is well established in image generation [119] and is being adapted to materials generation through reinforcement-learning guidance using stability or target-property rewards [120].

More broadly, these developments shift part of the upstream feasibility problem into the generator rather than relying entirely on post-hoc filtering. For example, several recent models embed symmetry directly into the generative process. CrystalFormer [121] generates space-group-controlled crystal structures by predicting Wyckoff positions, chemical elements, and fractional coordinates. SymmCD [122] models the asymmetric unit with the symmetry transformations for each atom in the diffusion process. LEGO-xtal [123] uses space-group and Wyckoff-based representations with local-environment targeting to build complex structures from only limited known examples.

This shift toward generation for feasibility is important because inverse design is inherently multimodal. A target performance profile is not translated into a deployable material solely through structure, but through the coupled selection of composition, processing route, microstructure, and expected response. In this setting, generative models need to do more than sample valid crystal structures. They also need to incorporate the conditions under which those structures can be realised and whether they remain compatible with downstream synthesis and performance constraints. Therefore, multimodal representations and knowledge integration provide the context needed to connect generated candidates to feasible material states rather than to abstract points in the design space.

Generative models are increasingly applied beyond structure proposal to synthesis itself. Diffusion models such as DiffSyn [124] generate synthesis routes, including compositions and synthesis conditions, conditioned on a target structure. LLMs have been used to propose inorganic synthesis conditions and precursors [125], and reinforcement learning with experimental feedback has been used to align reasoning traces with experimental outcomes [126].

These developments also change how generative models should be evaluated. Raw validity, diversity, or latent-space novelty are not sufficient if generated candidates fail under stability screening, violate symmetry or synthesis constraints, or cannot be connected to realistic processing routes. Therefore, emerging benchmarks increasingly evaluate trade-offs among validity, novelty, diversity, and constraint satisfaction under more realistic discovery settings [38]. This motivates the discussion in Sec. 5, where we examine benchmarking criteria that prioritise deployability over single-metric optimisation.

5 Evaluation and benchmarking

Benchmarks make models testable, but only when their reference data reflect the real-world discovery setting they are designed to evaluate. In AI for materials, this alignment remains incomplete. This section shows that many tasks rely on computational labels, that generative evaluation often depends on proxy metrics, and that experimental utility is rarely measured directly.

5.1 Evaluation and benchmarking landscape

Evaluation of materials AI models typically combines task definition, data splitting, and a set of performance metrics. Property prediction tasks are often framed as regression or classification, with error-based,

Table 2: Representative benchmarks for AI in materials research, grouped by task type and summarised by material domain, reference signal, evidence categories and data modalities. *Comp.*, *Microstruct.*, *Proc.* and *Test. & character.* denote the primary data categories. *Text*, *Graph*, *Image*, and *Spectra* denote the data modalities used to instantiate these data categories. The asterisk (*) indicates that multi-fidelity data are present, but paired settings are not provided, so multi-fidelity learning could be constructed when matched entries exist.

Benchmark	Year	Materials domain	Reference signal	Primary evidence categories					Primary data modalities			
				Comp.	Microstruct.	Proc.	Test. & character.	Multi-fidelity	Text	Graph	Image	Spectra
A. Property prediction												
Matbench Discovery [129]	2025	Crystals	DFT	✓	✓						✓	
TextEdge [96]	2025	Crystals	DFT	✓	✓				✓			
JARVIS-Leaderboard [130]	2024	Broad materials	DFT	✓	✓		✓	✓*	✓	✓	✓	✓
Matbench [131]	2020	Broad materials	Experimental	✓	✓			✓*		✓		
B. Generative modelling												
LeMat-GenBench [38]	2025	Crystals	Stability proxies	✓	✓						✓	
MGB [132]	2025	Crystals, MOFs	Proxy metrics	✓	✓						✓	
matbench-genmetrics [39]	2024	Crystals	Proxy metrics on time splits	✓	✓						✓	
C. Multimodal interpretation												
MATRIX [133]	2026	Materials reasoning	Prompts and rubrics	✓	✓	✓	✓		✓		✓	✓
MaCBench [134]	2025	Materials reasoning	Tasks and solutions	✓	✓	✓	✓		✓		✓	✓
MatQnA [135]	2025	Materials characterisation	QA pairs from papers				✓		✓		✓	✓

correlation-based, or threshold-based metrics used for comparison. Most studies still rely on random held-out splits or cross-validation, but these protocols can overestimate practical utility when test conditions differ from those seen during training. More recent evaluations use out-of-distribution settings that better reflect deployment conditions and evaluate the novelty of model predictions [24, 25, 26, 127]. In alloy systems, this can be framed in physically meaningful terms, such as whether a model trained on binary and ternary spaces can generalise to quaternary or higher-order alloys, or whether it can transfer across processing routes for nominally similar compositions. Such settings are more demanding than random splits, but they more closely reflect the challenges of real-world discovery.

Generative modelling requires a broader evaluation framework because outputs are evaluated both as samples from a learned distribution and as candidates for downstream screening. Common metrics include validity, novelty, uniqueness, and coverage, and recent toolkits adopt more consistent, leakage-resistant protocols, including time-aware splits for crystal datasets [39]. Deployment-oriented evaluation extends these criteria by including feasibility checks, such as chemical validity, structural consistency, and stability proxies under a chosen oracle, and by asking whether generated candidates satisfy the intended property or structural constraints rather than collapsing to easy-to-generate modes [128]. For conditional generation, useful metrics also include the success rate under target constraints and the best-of- N utility under a fixed screening budget, since practical workflows rank and filter batches rather than accept individual samples in isolation.

Benchmarks package these evaluations into a shared protocol by defining tasks, data splits, evaluation metrics, and baseline methods, often with public leaderboards. Table 2 summarises representative benchmarks by task type, materials domain, reference signals, primary input categories, and dominant modalities. Property-prediction benchmarks remain strongly shaped by crystalline materials with computational labels, where composition and crystal structure serve as the primary inputs and graph-based encodings are common. Suites such as Matbench [131], JARVIS-Leaderboard [130], and Matbench Discovery [129] have improved standardisation, while TextEdge [96] begins to extend crystal property prediction towards text-based inputs. Large-scale community benchmarks also exist for interatomic potentials and catalysis, such as the Open Catalyst Project [136, 137]. By contrast, generative benchmarks still rely heavily on proxy validation, particularly in inorganic crystal generation [38, 39, 132]. Emerging multimodal and LLM-based benchmarks increasingly probe cross-modal reasoning and interpretation of experimental artefacts [133, 134, 135], but process data remain under-represented relative to composition, structure, and observation.

5.2 Benchmarking challenges for deployability

Despite this growing benchmark ecosystem, several challenges remain. A fundamental challenge is the reliance on computational reference labels, such as DFT or related simulations. These benchmarks are valuable because they are scalable and internally consistent, but they only partially reflect experimental deployment. For example, for structural alloys, benchmark labels are often computed for idealised, defect-free unit cells at zero Kelvin, while real-world performance depends on finite-temperature phase stability, grain-boundary chemistry, dislocation structures, and thermomechanical history. Strong

performance on computational benchmarks does not necessarily translate into laboratory utility and may even reward shortcuts specific to the computational label distribution. The evaluation of generative models often depends on computational oracles or surrogate predictors, which introduce systematic bias and only partially reflect experimental feasibility [138, 139, 140].

Out-of-distribution evaluation also remains difficult to define in a physically meaningful and standardised way, making it insufficient to evaluate the novelty of model predictions. Although guided by domain knowledge, some out-of-distribution splits can still leave substantial overlap between training and test data in representation space. For example, this can happen when held-out alloy families share structural motifs and chemistry with the training set, or when test points fall within high-density regions of the compositional space already covered by training data. Subsequently, reported out-of-distribution performance reflects interpolation rather than true extrapolation, inflating claims of transferability [26].

Reporting practices remain inconsistent across benchmark studies. Computational cost, novelty-stability trade-offs, verification budgets, and success rates under constrained validation are not always reported in a comparable manner. Unified benchmark suites improve standardisation, but deployability will require evaluation protocols that connect accuracy and generative quality to experimentally relevant criteria, including feasibility, robustness, and performance under limited verification resources. Task-specific benchmark suites provide useful precedents, e.g., CSPBench [141], which shows that model rankings can depend strongly on test-set construction and structural similarity metrics.

The benchmarking challenges share a deeper common cause. Computational validation and experimental validity are not points on a single fidelity continuum but measures of fundamentally different things. Computational labels typically describe an idealised model system, such as a defect-free unit cell evaluated at zero Kelvin under a chosen exchange-correlation approximation. Experimental measurements instead reflect the behaviour of fabricated materials, which depends on finite temperature, defects, and processing history. Although these factors can be simulated, they usually fall outside the idealised setup used to generate benchmark labels. Therefore, the resulting computational and experimental targets are not simply low- and high-fidelity versions of the same quantity. Higher performance against computational benchmarks may not translate to experimental validity [44]. A model validated against computational reference signals has been tested only on its ability to reproduce a specific theoretical approximation of material behaviour. This provides limited evidence that the learned representations will transfer to laboratory conditions, because systematic errors in computational references can be correlated across training and test distributions in ways that experimental measurements are not [138, 139, 140].

These challenges show that benchmark performance is not necessarily correlated with deployable utility, while also clarifying the direction of progress. As evaluation moves beyond random splits, single-metric accuracy, and proxy-only validation, more realistic criteria are needed to evaluate whether generative and multimodal models can support the discovery of novel materials.

6 Future opportunities and recommendations

A consistent pattern emerges from the preceding sections. AI for materials research has made substantial progress on structural novelty, with generative models proposing compositionally and crystallographically diverse candidates at scale. However, progress slows markedly when moving up the hierarchy towards physical and deployment novelty. The remaining challenges are largely due to insufficient evidence grounding across the hierarchy. The evidence available to models is concentrated at the level of composition and idealised structure, whereas the evidence required to establish higher levels of novelty remains under-represented and weakly integrated.

The transition from structural to physical novelty is largely a data-coverage challenge with two key aspects. The first is the scarcity of processing and microstructural data: the modalities required to support physical novelty are far less available than compositional and structural data (Sec. 3.3). This scarcity reflects the economics of experimental materials science, where high-throughput screening or computation is relatively accessible, while detailed characterisation and long-duration testing remain costly and slow. Even when such modalities are reported, they are often coarse or poorly aligned with the composition and structure, limiting their value for attributing or comparing physical behaviour. The second is the scarcity of negative data, where “dark data” remains largely unreported under current success-oriented publication incentives. Without these counterexamples, models are susceptible to learning the interior of known compositional space, rather than the boundaries of what is physically realisable [84]. Generative models trained on such data can therefore support structural novelty but cannot establish whether a candidate is physically distinct from known materials.

The transition from physical to deployment novelty adds architectural and benchmarking challenges on top of these data challenges. Many multimodal models combine independently encoded modalities without learning the cross-modal interactions that govern deployment-relevant behaviour. Processing and microstructure, for example, are physically coupled: a given thermomechanical processing sequence

produces a specific grain boundary chemistry, and that link is a causal, physical one. A model that encodes processing and microstructure separately and simply concatenates the resulting embeddings may not adequately capture the relationship between them. Benchmarking poses a parallel challenge, as current benchmarks largely rely on computational proxies rather than experimental measurements, providing limited evidence of transfer to laboratory synthesis, manufacturing, or in-service conditions (Sec. 5.2).

A cross-cutting challenge underlying the three levels is a lack of standardisation. Data and metadata are reported in diverse, often incompatible formats, making it difficult to assemble consistently annotated multimodal datasets that span the full processing-structure-property chain. Without standardised schemas for consistent data reporting and an open infrastructure for data sharing, deployment novelty will remain poorly grounded, regardless of advances in AI models. Addressing these challenges requires community-wide infrastructure development.

6.1 Four opportunities for evidence-grounded discovery

The most immediate opportunity is to build standardised multimodal datasets that link each material to the context needed to interpret it. In such datasets, each material record should connect core data (composition, structure, microstructure, and properties) to the corresponding provenance metadata (e.g., synthesis route, processing conditions, measurement settings, units, uncertainties, and fidelity levels). This metadata should be captured and clearly reported at the source, because they are hard to recover afterwards, a key limitation of datasets assembled from the literature. To assemble such datasets, a standardised workflow is needed for generating and reporting experimental data, with shared schemas and reporting templates that keep modalities aligned, adapt across material classes and experiment types, and follow the FAIR principles [142]. Emerging closed-loop and self-driving laboratories, such as A-Lab [143] and AlphaFlow [144], provide an automated version of this workflow, acting not only as discovery platforms but also as generators of aligned multimodal data, including failed trials. Once assembled, these datasets can support multimodal data fusion and knowledge integration (Sec. 4.1 and Sec. 4.2).

Beyond constructing datasets, the next opportunity is to develop process-aware multimodal AI models that treat processing and manufacturing information as a primary modality rather than peripheral metadata. This can advance the model towards learning the full processing-structure-property relationship, rather than only mapping composition or crystal structure to a single target property. However, the processing modality is often scarce, coarse or inconsistently reported, requiring models to be robust to missing or heterogeneous inputs. Uncertainty-weighted fusion [145], for example, can down-weight unreliable inputs. Alongside statistical weighting, physical constraints can help compensate for missing data. Different modalities are often connected by underlying physical relationships, so a missing modality can sometimes be constrained or inferred using known physical laws. For example, thermodynamic phase rules can be used to partially infer missing microstructural information from a known composition and processing route. Such process-aware models are critical for predicting properties governed by cross-scale mechanisms that evolve over time, such as creep life, which is driven by the coupled evolution of dislocation motion, grain-boundary sliding and precipitate morphology, all shaped by processing history. Looking further ahead, these process-aware capabilities could be unified within materials-native multimodal foundation models pretrained on heterogeneous data and adapted to diverse downstream tasks, rather than being trained separately for each material class or task.

Building on process-aware representations, generative AI with a feasibility-first approach can propose candidates that are both high-performing and experimentally realisable. The premise is that the bottleneck in materials discovery is increasingly validation rather than candidate generation. In practice, this requires that physical, chemical, processing, and synthesis constraints be embedded in the representation, objective function, or sampling procedure, so that generated candidates remain within experimentally accessible regions. To evaluate feasibility, a generator should return a linked design tuple, comprising a candidate composition or structure, a feasible processing route, an expected microstructural state, target properties under defined testing conditions, and a calibrated uncertainty estimate. Generative models for synthesis planning [124, 125] are early steps in this direction. Generated candidates should then undergo staged experimental validation, progressing from computational screening and uncertainty-based ranking to small-batch synthesis and rapid characterisation, with long-duration tests reserved for the most promising candidates. Embedded in a closed-loop workflow, each round of synthesis and characterisation can feed back to refine the generator. This is a form of active learning: rather than testing candidates at random, the workflow prioritises experiments expected to be most informative, such as where the model is least certain, so that each test both filters candidates and improves the generator the most [146]. Validation should also extend beyond the as-processed state to service-relevant conditions, since deployment novelty depends on performance in service rather than initial properties. Multimodal AI can play an important role in service-life determination, linking processing history, exposure conditions and retained properties to estimate remaining useful life and rank candidates before full-duration testing [147].

In addition to validating candidates, AI models also need deployment-aware benchmarking that explicitly measures deployment novelty. This benchmarking should rest on three elements: out-of-distribution splits, deployment-relevant metrics, and calibrated uncertainty. Splits such as element holdout, compositional-complexity holdout, processing-route holdout, and time-aware tests aim to push models beyond the training data, but only if the held-out data are genuinely distant along relevant dimensions. Benchmarks should report this distance, so that extrapolation is not mistaken for interpolation. Evaluation metrics should reward deployment performance, such as constraint satisfaction, synthesis success rate, experimental confirmation rate and cost per confirmed candidate, rather than performance scores alone. Uncertainty estimates should be calibrated and tested across modalities, fidelity levels and out-of-distribution settings, using methods such as conformal prediction [148] or deep ensembles [149], so that they remain trustworthy precisely where models are most likely to fail. A deeper and still unresolved question underlies all of these protocols: when data are sparse, it is hard to tell whether a candidate is genuinely new or simply absent from a limited dataset, and no splitting strategy proposed to date fully resolves this.

These four opportunities, multimodal data construction, process-aware modelling, feasibility-first generation, and deployment-aware benchmarking, serve a common goal: grounding higher levels of novelty in stronger evidence. Achieving this requires not only technical progress but also collective action across the materials and AI communities, as discussed next.

6.2 Key recommendations

The opportunities above depend on data and the practices that govern how data are reported and shared, which calls for coordinated action across the community. A convening body with standing in the field should draft and maintain a versioned minimum metadata standard that specifies, at minimum, the composition, processing route, a microstructural descriptor, and the measurement conditions for each reported property. Journal editorial boards and funding bodies are well placed to enforce this standard. The former can require that metadata and negative results be deposited alongside reported compositions, much as structural biology made the deposition of atomic coordinates a condition of publication [150], albeit with the caveat that materials data are far more heterogeneous and resist a single fixed schema. The latter can attach the same requirements to data management plans, so that the cost of compliant reporting is carried by the grant. To avoid compliance becoming a burden for experimentalists, the community should also provide open, standardised infrastructure, including experimental workflow, metadata templates, data curation tools, and accessible deposition platforms such as GitHub, Hugging Face, or the UK Physical Sciences Data Infrastructure. Agentic AI can further reduce this burden by automating data collection and deposition, turning compliant reporting from a manual overhead into a background process.

Standards and infrastructure only have effect if those who produce and evaluate data act on them. Principal investigators running experimental and self-driving laboratories should report at the sample level, with each measurement linked to its processing and microstructural context, treat failed and null trials as first-class outputs, and, where possible, capture the exposure and post-exposure data needed for service-life prediction. Reviewers should treat the completeness of data and metadata as part of their assessment to ensure reporting standards are consistently enforced in practice. Benchmark maintainers should publish the reference sets and the comparability criteria used to ensure novelty claims to be compared across studies. Taken together, these recommendations provide the opportunities outlined in this section with the evidence they rely on.

7 Conclusion

This Perspective explored whether and how generative and multimodal AI can support the design of novel materials. We pointed out that novelty cannot be reduced to statistical distance in composition or latent space, and we introduced a three-level taxonomy of structural, physical and deployment novelty for evaluating AI-proposed candidates, together with a hierarchy that spans intrinsic properties to extrinsic performance. Seen through this framework, progress has been strong at the structural level, where models now propose diverse and stable candidates at scale, but weak at the physical and deployment novelty levels. The reason is a shortage of evidence: the processing, microstructural and experimental data needed to establish higher-level novelty are scarce, poorly integrated, and evaluated largely against computational proxies. Closing this gap needs a collective undertaking. The direction we have outlined points towards systems that return not a bare composition or structure, but a complete and experimentally grounded design, one whose feasibility can be tested rather than assumed. Achieving it will require that the data, models, generation, and validation steps be developed together rather than in isolation, and that the evidence on which they depend becomes a shared responsibility of the materials and AI communities. Whether AI can design novel materials, rather than only rediscover known ones, will depend not only on model capacity but also on the evidence we work together to build.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Acknowledgments

X.L., C.A., N.A.M., and H.L. acknowledge funding from EPSRC (grant UKRI396). K.A.C, and B.E.J. acknowledge funding from EP/R00661X/1 and EP/S019367/1. A.J.R acknowledges funding from EPSRC (grant EP/X039285/1). We thank X. Yang and H. Hassan for collecting related papers.

Data availability

The source data for Figures 1 and 3 are available at <https://github.com/multimodalAI/multimodal-ai-landscape/tree/main/data/material-landscape> and will be updated annually. An interactive explorer is also available at <https://multimodalai.github.io/multimodal-ai-landscape/material.html> to facilitate navigation.

References

[1] Merchant A, Batzner S, Schoenholz S S, Aykol M, Cheon G and Cubuk E D 2023 *Nature* **624** 80–85

[2] Zeni C, Pinsler R, Züchner D, Fowler A, Horton M, Fu X, Wang Z, Shysheya A, Crabbé J, Ueda S *et al.* 2025 *Nature* **639** 624–632

[3] Hashemi S M, Parvizi S, Baghbanijavid H, Tan A T, Nematollahi M, Ramazani A, Fang N X and Elahinia M 2022 *International Materials Reviews* **67** 1–46

[4] Kang Y, Park H, Smit B and Kim J 2023 *Nature Machine Intelligence* **5** 309–318

[5] Moro V, Loh C, Dangovski R, Ghorashi A, Ma A, Chen Z, Kim S, Lu P Y, Christensen T and Soljačić M 2025 *Newton* **1**

[6] Wu Y, Ding M, He H, Wu Q, Jiang S, Zhang P and Ji J 2025 *npj Computational Materials* **11**

[7] Joshi C K, Fu X, Liao Y L, Gharakhanyan V, Miller B K, Sriram A and Ulissi Z W 2025 All-atom diffusion transformers: Unified generative modelling of molecules and materials *International Conference on Machine Learning*

[8] Ye C, Wang Y, Xie X, Zhu T, Liu J, He Y, Zhang L, Zhang J, Fang Z, Wang L *et al.* 2025 *npj Computational Materials* **12**

[9] Chenebueh E T, Nganbe M and Tchagang A B 2024 *npj Computational Materials* **10**

[10] Jiao R, Huang W, Liu Y, Zhao D and Liu Y 2024 Space group constrained crystal generation *International Conference on Learning Representations*

[11] Xie T and Grossman J C 2018 *Physical Review Letters* **120** 145301

[12] Choudhary K and DeCost B 2021 *npj Computational Materials* **7**

[13] Deng B, Zhong P, Jun K, Riebesell J, Han K, Bartel C J and Ceder G 2023 *Nature Machine Intelligence* **5** 1031–1041

[14] Ye C Y, Weng H M and Wu Q S 2024 *Computational Materials Today* **1** 100003

[15] Luo X, Wang Z, Gao P, Lv J, Wang Y, Chen C and Ma Y 2024 *npj Computational Materials* **10**

[16] Khajeh A, Lei X, Ye W, Yang Z, Hung L, Schweigert D and Kwon H K 2025 *Digital Discovery* **4** 11–20

[17] Lambard G, Yamazaki K and Demura M 2023 *Scientific Reports* **13** 566

[18] Altoyuri A H, Sarmah A and Jain M K 2024 *Acta Materialia* **281** 120431

[19] Jiang X, Wang W, Tian S, Wang H, Lookman T and Su Y 2025 *npj Computational Materials* **11**

[20] Schilling-Wilhelmi M, Ríos-García M, Shabih S, Gil M V, Miret S, Koch C T, Márquez J A and Jablonka K M 2025 *Chemical Society Reviews* **54** 1125–1150

[21] Liu X, Ouyang B and Zeng Y 2025 *Nature Computational Science* **5** 92–94

[22] Zhang L, Bing Q, Qin H, Yu L, Li H and Deng D 2025 *Matter* **8**

[23] Liu X, Zhang J, Zhou S, van der Plas T L, Vijayaraghavan A, Grishina A, Zhuang M, Schofield D, Tomlinson C, Wang Y *et al.* 2025 *Nature Machine Intelligence* 1–13

- [24] Jablonka K M, Rosen A S, Krishnapriyan A S and Smit B 2023 *ACS Central Science* **9** 563–581
- [25] Omee S S, Fu N, Dong R, Hu M and Hu J 2024 *npj Computational Materials* **10**
- [26] Li K, Rubungo A N, Lei X, Persaud D, Choudhary K, DeCost B, Dieng A B and Hatrnick-Simpers J 2025 *Communications Materials* **6** 9
- [27] Jain A, Ong S P, Hautier G, Chen W, Richards W D, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G *et al.* 2013 *APL Materials* **1**
- [28] Kirklin S, Saal J E, Meredig B, Thompson A, Doak J W, Aykol M, Rühl S and Wolverton C 2015 *npj Computational Materials* **1**
- [29] Saal J E, Kirklin S, Aykol M, Meredig B and Wolverton C 2013 *Jom* **65** 1501–1509
- [30] Choudhary K, Garrity K F, Reid A C, DeCost B, Biacchi A J, Hight Walker A R, Trautt Z, Hatrnick-Simpers J, Kusne A G, Centrone A *et al.* 2020 *npj Computational Materials* **6**
- [31] Antoniuk E R, Cheon G, Wang G *et al.* 2023 *npj Computational Materials* **155**
- [32] Walsh A and Zunger A 2017 *Nature Materials* **16** 964–967
- [33] Xie J, Zhou Y, Faizan M, Li Z, Li T, Fu Y, Wang X and Zhang L 2024 *Nature Computational Science* **4** 322–333
- [34] Zheng Y, Xu H, Li Z, Li L, Yu Y, Jiang P, Shi Y, Zhang J, Huang Y, Luo Q, Lou Z and Wang L 2025 *Advanced Materials* **37** 2504378
- [35] Yang H, Hu C, Zhou Y, Liu X, Shi Y, Li J, Li G, Chen Z, Chen S, Zeni C *et al.* 2024 *arXiv preprint arXiv:2405.04967*
- [36] Hayes S M and McCullough E A 2018 *Resources Policy* **59** 192–199
- [37] Mannan S, Myers R J, Batra R, Mercado R, Wondraczek L and Krishnan N 2026 *arXiv preprint arXiv:2601.21527*
- [38] Betala S, Gleason S P, Ramlaoui A, Xu A, Channing G, Levy D, Fourrier C, Kazeev N, Joshi C K, Kaba S O *et al.* 2025 *arXiv preprint arXiv:2512.04562*
- [39] Baird S G, Sayeed H M, Montoya J and Sparks T D 2024 *Journal of Open Source Software* **9** 5618
- [40] Gruver N, Sriram A, Madotto A, Wilson A G, Zitnick L and Ulissi Z 2024 Fine-tuned language models generate stable inorganic materials as text *International Conference on Learning Representations*
- [41] Ong S P, Richards W D, Jain A, Hautier G, Kocher M, Cholia S, Gunter D, Chevrier V L, Persson K A and Ceder G 2013 *Computational Materials Science* **68** 314–319
- [42] Widdowson D E and Kurlin V A 2026 *SIAM Journal on Applied Mathematics* **86** 898–918
- [43] Juelsholt M 2026 *Materials Horizons* **13**(11) 5672–5679
- [44] Cheetham A K and Seshadri R 2024 *Chemistry of Materials* **36** 3490–3495
- [45] Yu D, Griesemer S, Liu T c, Wolverton C and Zhu Y 2025 *npj Computational Materials* **11**
- [46] Meredig B, Agrawal A, Kirklin S, Saal J E, Doak J W, Thompson A, Zhang K, Choudhary A and Wolverton C 2014 *Physical Review B* **89** 094104
- [47] Ward L, Agrawal A, Choudhary A and Wolverton C 2016 *npj Computational Materials* **2**
- [48] Butler K T, Davies D W, Cartwright H, Isayev O and Walsh A 2018 *Nature* **559** 547–555
- [49] Wang Z, Wang L, Zhang H, Xu H and He X 2024 *Nano Convergence* **11** 8
- [50] Kononova O, Huo H, He T, Rong Z, Botari T, Sun W, Tshitoyan V and Ceder G 2019 *Scientific Data* **6** 203
- [51] Lee S, Cruse K, Baibakova V, Ceder G and Jain A 2025 *Scientific Data*
- [52] Teng Y, Tan H, Huang W and Shan G 2025 *Communications Physics* **8**

- [53] Wissel S, Scheunert J, Dextre A, Ahmed S, Beyer A, Volz K and Xu B X 2025 *Materials & Design* **115**069
- [54] Wanni J, Bronkhorst C A and J T D 2024 *npj Computational Materials* **10**
- [55] Ling J, Hutchinson M, Antono E, DeCost B, Holm E A and Meredig B 2017 *Materials Discovery* **10** 19–28
- [56] Tsuruta H and Kumagai M 2025 MatPROV: A provenance graph dataset of material synthesis extracted from scientific literature *NeurIPS 2025 Workshop on AI for Accelerated Materials Design*
- [57] Kim H, Jeon T, Choi S, Hong J H, Jeon D W, Baek G Y, Kwak G W, Lee D H, Bae J, Lee C *et al.* 2025 Towards fully-automated materials discovery via large-scale synthesis dataset and expert-level llm-as-a-judge *ACM International Conference on Information and Knowledge Management* pp 1302–1312
- [58] Zhang J, Yin C, Farbiz F, Jafary-Zadeh M and Sing S L 2025 *Virtual and Physical Prototyping* **20** e2592732
- [59] Deshmukh K, Riensche A, Bevans B, Lane R J, Snyder K, Halliday H S, Williams C B, Mirzaeifar R and Rao P 2024 *Materials & Design* **244** 113136
- [60] Lederbauer M, Betala S, LI X, Jain A, Sehaba M E A, Channing G, Germain G, Leonescu A, Flaifil F, Amayuelas A, Nozadze A, Schmid S P, Zaki M, Ethirajan S K, Pan E, Franckel M L D, Duval A, Krishnan N M A and Gleason S P 2025 Lemat-synth: a multi-modal toolbox to curate broad synthesis procedure databases from scientific literature *NeurIPS 2025 Workshop on AI for Accelerated Materials Design*
- [61] Salgado J E, Lerman S, Du Z, Xu C and Abdolrahim N 2023 *npj Computational Materials* **9**
- [62] Oviedo F, Ren Z, Sun S, Settens C, Liu Z, Hartono N T P, Ramasamy S, DeCost B L, Tian S I, Romano G *et al.* 2019 *npj Computational Materials* **5**
- [63] Paulus B and Biskup T 2023 *Digital Discovery* **2** 234–244
- [64] Statt M J, Rohr B A, Guevarra D, Suram S K, Morrell T E and Gregoire J M 2023 *Scientific Data* **10** 184
- [65] Swain M C and Cole J M 2016 *Journal of Chemical Information and Modeling* **56** 1894–1904
- [66] Mavracic J, Court C J, Isazawa T, Elliott S R and Cole J M 2021 *Journal of Chemical Information and Modeling* **61** 4280–4289
- [67] Polak M P and Morgan D 2024 *Nature Communications* **15** 1569
- [68] Dagdelen J, Dunn A, Lee S, Walker N, Rosen A S, Ceder G, Persson K A and Jain A 2024 *Nature Communications* **15** 1418
- [69] Hira K, Zaki M, Krishnan N *et al.* 2025 *arXiv preprint arXiv:2509.10448*
- [70] Li C, Xian Y, Zhou Y, Ding X, Sun J and Xue D 2026 *Advanced Materials* e20478
- [71] Debnath A, Krajewski A M, Sun H, Lin S, Ahn M, Li W, Priya S, Singh J, Shang S, Beese A M, Liu Z K and Reinhart W F 2021 *Journal of Materials Informatics* **1** 3
- [72] Reeves-McLaren N and Christensen S M L 2026 *Journal of Materials Chemistry A* **14** 276–283
- [73] Berry J and Christofidou K A 2025 *Materials Science and Technology* **41** 773–790
- [74] Wang A Y T, Murdock R J, Kauwe S K, Oliynyk A O, Gurlo A, Brgoch J, Persson K A and Sparks T D 2020 *Chemistry of Materials* **32** 4954–4965
- [75] Chang R, Wang Y X and Ertekin E 2022 *npj Computational Materials* **8**
- [76] Xu P, Ji X, Li M and Lu W 2023 *npj Computational Materials* **9**
- [77] Che L, He Z, Zheng K, Si T, Ge M, Cheng H and Zeng L 2023 *Materials Today Communications* **37** 107531

- [78] Wong R, Tran A, Dovggy B, Maldonado C S and Pham M S 2025 *Materials & Design* **256** 114301
- [79] Gianola D S, della Ventura N M, Balbus G H, Ziemke P, Echlin M P and Begley M R 2023 *Current Opinion in Solid State and Materials Science* **27** 101090
- [80] Mouritz A P 2012 *Introduction to aerospace materials* (Elsevier) ISBN 0857095153
- [81] Horton M K, Huck P, Yang R X, Munro J M, Dwaraknath S, Ganose A M, Kingsbury R S, Wen M, Shen J X, Mathis T S et al. 2025 *Nature Materials* **24** 1522–1532
- [82] Liu Y, Wang C, Liu J, Shi X, Huang Y, Cheng Q and Lu W 2026 *Scientific Data* **13**
- [83] Boyce B, Dingreville R, Desai S, Walker E, Shilt T, Bassett K L, Wixom R R, Stebner A P, Arroyave R, Hattrick-Simpers J and Warren J A 2023 *Matter* **6** 1320–1323
- [84] Maloney M P, Coley C W, Genheden S, Carson N, Helquist P, Norrby P O and Wiest O 2023 *Organic Letters* **25** 2945–2947
- [85] Kozlowski M C 2022 *Organic Letters* **24** 7247–7249
- [86] Himanen L, Geurts A, Foster A S and Rinke P 2019 *Advanced Science* **6** 1900808
- [87] Goodall R E and Lee A A 2020 *Nature Communications* **11** 6280
- [88] Wang A Y T, Kauwe S K, Murdock R J and Sparks T D 2021 *Npj Computational Materials* **7**
- [89] Jha D, Ward L, Paul A, Liao W k, Choudhary A, Wolverton C and Agrawal A 2018 *Scientific Reports* **8** 17593
- [90] Kim E, Huang K, Saunders A, McCallum A, Ceder G and Olivetti E 2017 *Chemistry of Materials* **29** 9436–9444
- [91] Wang Z, Kononova O, Cruse K, He T, Huo H, Fei Y, Zeng Y, Sun Y, Cai Z, Sun W et al. 2022 *Scientific Data* **9** 231
- [92] Vecsei P M, Choo K, Chang J and Neupert T 2019 *Physical Review B* **99** 245120
- [93] Simonnet T, Grangeon S, Claret F, Maubec N, Fall M D, Harba R and Galerne B 2024 *IUCrJ* **11**
- [94] Gupta T, Zaki M, Krishnan N A and Mausam 2022 *npj Computational Materials* **8**
- [95] Shetty P, Rajan A C, Kuenneth C, Gupta S, Panchumarti L P, Holm L, Zhang C and Ramprasad R 2023 *npj Computational Materials* **9**
- [96] Niyongabo Rubungo A, Arnold C, Rand B P and Dieng A B 2025 *npj Computational Materials* **11**
- [97] Jiang T, Xiong J, Su T, Chen S and Zhang T Y 2026 *Journal of Physics: Materials* **9** 025003
- [98] Venugopal V and Olivetti E 2024 *Scientific Data* **11** 217
- [99] Zhang Y, Chen F, Liu Z, Ju Y, Cui D, Zhu J, Jiang X, Guo X, He J, Zhang L et al. 2024 *Scientific Data* **11** 600
- [100] Ye Y, Ren J, Wang S, Wan Y, Razzak I, Hoex B, Wang H, Xie T and Zhang W 2024 *Advances in Neural Information Processing Systems* **37** 56878–56897
- [101] Song G, Fu D, Qiu Z, Yang Z, Dai J, Ma L and Zhang D 2025 *Applied Soft Computing* 113776
- [102] Mrdjenovich D, Horton M K, Montoya J H, Legaspi C M, Dwaraknath S, Tshitoyan V, Jain A and Persson K A 2020 *Matter* **2** 464–480
- [103] Bai X, He S, Li Y, Xie Y, Zhang X, Du W and Li J R 2025 *npj Computational Materials* **11**
- [104] Takeda S, Kishimoto A, Hamada L, Nakano D and Smith J R 2023 Foundation model for material science *Proceedings of the AAAI Conference on Artificial Intelligence* vol 37 pp 15376–15383
- [105] Fu N, Wei L, Song Y, Li Q, Xin R, Omee S S, Dong R, Siriwardane E M D and Hu J 2023 *Machine Learning: Science and Technology* **4** 015001

- [106] Zhang Y, Chen X, Jin B, Wang S, Ji S, Wang W and Han J 2024 A comprehensive survey of scientific large language models and their applications in scientific discovery *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* pp 8783–8817
- [107] Zhang J, Zhang D C, Liang S, Li Z, Ying R and Shao J 2024 Retrieval-augmented language model for knowledge-aware protein encoding *International Conference on Machine Learning*
- [108] Maharana P R, Verma A and Joshi K 2025 *Journal of Physics: Materials* **8** 035006
- [109] Manakul P, Liusie A and Gales M 2023 Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* pp 9004–9017
- [110] Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, Peng W, Feng X, Qin B *et al.* 2025 *ACM Transactions on Information Systems* **43** 1–55
- [111] Lee J, Wang Y, Li J and Zhang M 2024 Multimodal reasoning with multimodal knowledge graph *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* pp 10767–10782
- [112] Huang Z, Li Z, Jiang H, Cao T, Lu H, Yin B, Subbian K, Sun Y and Wang W 2022 Multilingual knowledge graph completion with self-supervised adaptive graph alignment *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* pp 474–485
- [113] Cheng M, Fu C L, Okabe R, Chotrattanapituk A, Boonkird A, Hung N T and Li M 2026 *Nature Materials* 1–17
- [114] Xie T, Fu X, Ganea O E, Barzilay R and Jaakkola T 2022 Crystal diffusion variational autoencoder for periodic material generation *International Conference on Learning Representations*
- [115] Alverson M, Baird S G, Murdock R, Ho E S H, Johnson J and Sparks T D 2024 *Digital Discovery* **3** 62–80
- [116] Mohanty T, Mehta M, Sayeed H M, Oded B E, Pitussi I, Borenstein A, Srikumar V and Sparks T D 2026 *Integrating Materials and Manufacturing Innovation* **15** 416–430
- [117] Okabe R, Cheng M, Chotrattanapituk A, Mandal M, Mak K, Córdova Carrizales D, Hung N T, Fu X, Han B, Wang Y, Xie W, Cava R J, Jaakkola T S, Cheng Y, Li M *et al.* 2026 *Nature Materials* **25** 223–230
- [118] Reents T, Cantarella A, Bercx M, Bonfà P and Pizzi G 2026 *arXiv preprint arXiv:2601.01959*
- [119] Chung H, Kim J, McCann M T, Klasky M L and Ye J C 2023 Diffusion posterior sampling for general noisy inverse problems *International Conference on Learning Representations*
- [120] Park H and Walsh A 2025 *arXiv preprint arXiv:2511.07158*
- [121] Cao Z, Luo X, Lv J and Wang L 2025 *Science Bulletin* **70** 3522–3533
- [122] Levy D, Panigrahi S S, Kaba S O, Zhu Q, Galkin M, Miret S and Ravanbakhsh S 2024 SymmCD: Symmetry-preserving crystal generation with diffusion models *NeurIPS 2024 Workshop on AI for Accelerated Materials Design*
- [123] Ridwan O G, Pitié S, Raj M S, Dai D, Frapper G, Xue H and Zhu Q 2026 *npj Computational Materials* **12**
- [124] Pan E, Kwon S, Liu S, Xie M, Hoffman A J, Duan Y, Prein T, Sherif K, Roman-Leshkov Y, Moliner M *et al.* 2026 *Nature Computational Science* **6** 404–416
- [125] Prein T, Pan E, Jehkul J, Weinmann S, Olivetti E and Rupp J L 2025 *ACS Applied Materials & Interfaces* **17** 69221–69233
- [126] Pan E, Prein T, Nam J, Du X, Yang S, Cai P, Rupp J L, Gomez-Bombarelli R and Olivetti E 2026 Synreason: Enhancing synthesis reasoning via reinforcement learning experimental feedback *ICLR 2026 Workshop on AI for Accelerated Materials Design*
- [127] Witman M D and Schindler P 2025 *Digital Discovery* **4** 625–635

- [128] De Breuck P P, Wang H C, Rignanese G M, Botti S and Marques M A L 2025 *npj Computational Materials* **11**
- [129] Riebesell J, Goodall R E, Benner P, Chiang Y, Deng B, Ceder G, Asta M, Lee A A, Jain A and Persson K A 2025 *Nature Machine Intelligence* **7** 836–847
- [130] Choudhary K, Wines D, Li K, Garrity K F, Gupta V, Romero A H, Krogel J T, Saritas K, Fuhr A, Ganesh P et al. 2024 *npj Computational Materials* **10**
- [131] Dunn A, Wang Q, Ganose A, Dopp D and Jain A 2020 *npj Computational Materials* **6**
- [132] Yan L, Kang B S, Hanisch M D, Ma J and Anandkumar A Mgb: The material generation benchmark *NeurIPS 2025 Workshop on AI for Accelerated Materials Design*
- [133] McGrath D, Chong C, Kulkarni R, Ceder G and Kolluru A 2026 *arXiv preprint arXiv:2602.00376*
- [134] Alampara N, Schilling-Wilhelmi M, Ríos-García M, Mandal I, Khetarpal P, Grover H S, Krishnan N A and Jablonka K M 2025 *Nature Computational Science* **5** 952–961
- [135] Weng Y, Gao L, Zhu L and Huang J 2025 *arXiv preprint arXiv:2509.11335*
- [136] Chanussot L, Das A, Goyal S, Lavril T, Shuaibi M, Riviere M, Tran K, Heras-Domingo J, Ho C, Hu W et al. 2021 *Acs Catalysis* **11** 6059–6072
- [137] Tran R, Lan J, Shuaibi M, Wood B M, Goyal S, Das A, Heras-Domingo J, Kolluru A, Rizvi A, Shoghi N et al. 2023 *Acs Catalysis* **13** 3066–3084
- [138] Perdew J P 1985 *International Journal of Quantum Chemistry* **28** 497–523
- [139] Harvey J N 2006 *Annual Reports Section C (Physical Chemistry)* **102** 203–226
- [140] Lejaeghere K, Van Speybroeck V, Van Oost G and Cottenier S 2014 *Critical Reviews in Solid State and Materials Sciences* **39** 1–24
- [141] Wei L, Ome S S, Dong R, Fu N, Song Y, Siriwardane E, Xu M, Wolverton C and Hu J 2024 *arXiv preprint arXiv:2407.00733*
- [142] Wilkinson M D, Dumontier M, Aalbersberg I J, Appleton G, Axton M, Baak A, Blomberg N, Boiten J W, da Silva Santos L B, Bourne P E et al. 2016 *Scientific Data* **3**
- [143] Szymanski N J, Rendy B, Fei Y, Kumar R E, He T, Milsted D, McDermott M J, Gallant M, Cubuk E D, Merchant A et al. 2023 *Nature* **624** 86–91
- [144] Volk A A, Epps R W, Yonemoto D T, Masters B S, Castellano F N, Reyes K G and Abolhasani M 2023 *Nature Communications* **14**
- [145] Han Z, Zhang C, Fu H and Zhou J T 2022 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45** 2551–2566
- [146] Lookman T, Balachandran P V, Xue D and Yuan R 2019 *npj Computational Materials* **5**
- [147] Wang H, Li B, Gong J and Xuan F Z 2023 *Engineering Fracture Mechanics* **284**
- [148] Angelopoulos A N and Bates S 2023 *Foundations and Trends in Machine Learning* **16** 494–591
- [149] Lakshminarayanan B, Pritzel A and Blundell C 2017 *Advances in Neural Information Processing Systems* **30**
- [150] Berman H, Henrick K and Nakamura H 2003 *Nature Structural & Molecular Biology* **10**