LEMAT-GENBENCH: Bridging the gap between crystal generation and materials discovery

Alexandre Duval
Entalpic
`alexandre.duval@entalpic.ai`

Siddharth Betala
Entalpic
`siddharth.betala-ext@entalpic.ai`

Samuel P. Gleason
Entalpic

Andy Xu
Harvey Mudd College

Georgia Channing
Hugging Face

Daniel Levy
Mila

Ali Ramlaoui
Entalpic

Clementine Fourrier
Hugging Face

Chaitanya K. Joshi
University of Cambridge

Nikita Kazeev
National University of Singapore

Sekou-Oumar Kaba
Mila

Felix Therrien
Mila

Alex Hernandez-Garcia
Mila

Rocio Mercado
Chalmers University of Technology

N. M. Anoop Krishnan
IIT Delhi

Abstract

Generative machine learning models hold great promise for accelerating materials discovery, particularly through the inverse design of inorganic crystals—enabling an unprecedented exploration of chemical space. Yet, the lack of standardized benchmarks makes it difficult to evaluate, compare and further develop these ML models meaningfully. In this benchmark paper, we introduce LEMAT-GENBENCH, a unified framework for assessing generative models of crystalline materials. In particular, we propose a set of evaluation metrics alongside a set of tasks (unconditional, conditional, and limited-budget crystal generation), designed to better inform model developers as well as downstream practical applications. To support it, we release an open-source evaluation suite and a public leaderboard on Hugging Face with verified submissions. Altogether, LEMAT-GENBENCH aims to guide model development and bridge the gap between generative modeling and practical materials discovery.

1 Introduction

The discovery process of inorganic crystalline materials has traditionally relied on an Edisonian cycle of expert intuition and experimental validation [Schmidt et al., 2019], occasionally aided by quantum simulations such as density functional theory (DFT) [Kohn et al., 1996, Sholl and Steckel, 2009]. While first-principle simulations provide valuable insights into structure and stability, they remain computationally expensive and require predefined atomic configurations, which are non-trivial to generate for novel crystal structures. Machine Learning (ML) models [Deringer et al., 2019, Unke et al., 2021], particularly those based on geometric graph neural networks [Duval et al., 2023a], offer faster alternatives to DFT, enabling scalable evaluation of candidate structures. Yet, like DFT, they are limited to assessing pre-defined inputs and are not designed to explore and propose novel materials.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI4Mat-NeurIPS-2025: NeurIPS 2025 Workshop on AI for Accelerated Materials Design.

This has motivated a growing wave of generative ML models for materials discovery [Zeni et al., 2023, Mila AI4Science et al., 2023, Levy et al., 2025a, Kazeev et al., 2025]. These models—ranging from variational autoencoders (VAEs) [Kingma and Welling, 2013] and diffusion models [Song et al., 2021] to GFlowNets [Bengio et al., 2023] and large language models (LLMs) [Brown et al., 2020]—aim to learn the distribution of valid crystal structures and sample from it, ideally guided by target properties. This *inverse design* paradigm promises to unlock previously inaccessible regions of chemical space and accelerate the discovery of practically relevant materials.

The rapid development of generative models for crystal structures has revealed a critical challenge: the absence of standardized evaluation protocols rooted in applications. Studies vary widely in how they assess stability, define novelty, or validate structures; using different reference datasets, fingerprinting methods, energy estimators [Batatia et al., 2023, Unke et al., 2021], relaxation procedures, energy thresholds, etc. Without shared benchmarks, it remains difficult to disentangle genuine model improvements from differences in evaluation design. While standardized benchmarks like Matbench [Dunn et al., 2020] have transformed property prediction by enabling systematic model comparison, no analogous infrastructure exists for the evaluation of generative models of crystal structures. This gap makes progress difficult to quantify, comparisons unfair, and leaves the community without shared reference standards for scientific advancement.

To this end, we introduce LEMAT-GENBENCH, a benchmarking framework aimed at standardizing the evaluation of generative models of inorganic crystal structures. LEMAT-GENBENCH provides:

- **Standardized evaluation metrics:** A unified suite centered on the (conditional) (Meta)Stable, Unique, Novel (M.S.U.N.) rate, plus validity, diversity, efficiency and multi-objective metrics;
- **Diverse evaluation tasks:** Benchmarks spanning unconditional, property-conditioned, and limited-budget generation, designed to better reflect practical downstream discovery scenarios;
- **Open evaluation infrastructure:** A public Python toolkit and Hugging Face leaderboard with evaluations of 10 contemporary generative models.

By establishing shared protocols and rigorous evaluation standards, LEMAT-GENBENCH aims to enable more systematic, fair, and transparent model comparisons. The framework is designed to evolve with advancing methodologies while supporting both research and practical applications in AI-driven materials.

2 The State of Generative Modeling for Crystal Structures

The field of generative modeling for inorganic crystals has rapidly diversified, with a growing number of architectures, representations, and conditioning strategies being proposed. A taxonomy is illustrated in Figure 2. While these methods share the goal of automating candidate generation for materials discovery, they differ significantly in how they represent crystals, enforce physical constraints, and guide sampling toward desired properties. In what follows, we provide a structured overview of current modeling approaches (Section 2.1), followed by an analysis of existing evaluation practices and their limitations (Section 2.2). Together, these point to the need for standardized benchmarks and motivate the framework we introduce in this work.

2.1 Modeling Approaches Overview

Generative modeling for inorganic crystals has emerged as a promising strategy for inverse design, enabling the proposal of candidate structures with desired stability, symmetry, or functional properties. These ML models are typically trained on large datasets of relaxed crystal structures and generate new candidates either unconditionally or conditioned on specific targets. Over the past few years, a range of architectural paradigms have been explored. Figure 2 provides a visual illustration. We briefly discuss the prominent family of models below.

Latent-variable models such as variational autoencoders (VAEs) were among the earliest approaches [Noh et al., 2019, Hoffmann et al., 2019, Court et al., 2020]. These methods encode crystal representations into a continuous latent space to enable interpolation and sampling, but struggle with decoding to valid atomic configurations, especially when relying on voxel grids or density maps [Hoffmann et al., 2019, Zhao et al., 2023].

GAN-based methods attempt to generate crystal fingerprints or coordinate-based representations via adversarial training [Kim et al., 2020, Zhao et al., 2021]. While conceptually appealing, these models suffer from training instability and limited diversity, and typically do not scale well to 3D periodic systems Zhao et al. [2021].

Diffusion models have become the dominant paradigm, leveraging score-based denoising to gradually transform noise into periodic atomic structures [Xie et al., 2021, Jiao et al., 2023, Zeni et al., 2025]. By jointly modeling atom types, fractional coordinates, and lattice parameters—often with SE(3)-equivariant or symmetry-aware networks—these models learn to generate valid and plausible crystals. Recent variants incorporate explicit symmetry conditioning [Jiao et al., 2024, Levy et al., 2025a, Chang et al., 2025] or textual guidance from pretrained language models [Das et al., 2025, Park et al., 2025].

Flow-based models offer a related but computationally faster alternative. Flow matching methods learn continuous velocity fields to map between base distributions and the data manifold in one pass [Lipman et al., 2023]. Crystal-specific variants such as FlowMM [Miller et al., 2024] and FlowLLM [Sriram et al., 2024] combine geometric inductive biases with learned base distributions informed by large language models [Gruver et al., 2024]. ADiT [Joshi et al., 2025] further generalizes this approach to generate both crystals and molecules in a shared latent space. Puny et al. [2025] developed a space group equivariant flow matching method.

Sequential generation strategies like reinforcement learning (RL) [Zamaraeva et al., 2023, Govindarajan et al., 2024] and Generative Flow Networks (GFlowNets) [Mila AI4Science et al., 2023, Cipcigan et al., 2024, Podina et al., 2025] decompose generation into stepwise actions guided by a reward. These methods allow the introduction of hard constraints, flexible conditioning (e.g., on Wyckoff positions or energy-based rewards) and encourage diverse generation through reward-proportional sampling.

Large language models (LLMs) have recently been adapted to crystal generation by tokenizing CIF formats or leveraging textual prompts. Models such as CrystalLLM [Gruver et al., 2024], WyFormer [Kazeev et al., 2025], and PLaID [Xu et al., 2025a] have demonstrated strong performance using autoregressive decoding and reinforcement learning. Some also enable multimodal conditioning (e.g., with PXRD or synthesis data) [Johansen et al., 2025, Moro et al., 2025], suggesting new directions for generative systems grounded in experimental context.

Taken together, these methods reflect an increasingly diverse and dynamic modeling landscape. However, despite architectural progress, evaluating and comparing generative models remains a major open challenge—due to the lack of standardized tasks, metrics, and data splits.

2.2 Evaluation Metrics and Benchmarking Efforts

Evaluating generative models for crystal structures is inherently more complicated than supervised tasks like property prediction, where objective metrics such as RMSE or MAE are well established. It requires assessing both structural plausibility and functional utility without ground-truth references, leading to wide variation in evaluation protocols and limited comparability across works.

Most studies focus on generating structures that are **valid** and **stable**. Validity is typically assessed via heuristic filters such as charge neutrality, interatomic distances, or CIF readability [Xie et al., 2021]. However, these checks vary across papers, serve different purposes (either as hard pre-filters or as indicative sanity checks) and are not always consistent with real-world data. For example, many Materials Project entries fail certain validity tests. Stability is usually evaluated through formation energy (E_f) and energy above the convex hull (E_{hull}), either via DFT or machine-learned interatomic potentials (MLIPs) [Batatia et al., 2023]. However, threshold choices (e.g., $E_{\text{hull}} < 0$ vs. < 0.1 eV/atom), relaxation strategies, and reference datasets vary, making comparisons difficult. MLIPs scale well but tend to underestimate E_f [Fu et al., 2022, Nong et al., 2025], and few works quantify this uncertainty or reconcile MLIP-based scores with DFT-based hulls.

Building upon stability, the **S.U.N.** framework (Stable, Unique, Novel) [Zeni et al., 2025] has emerged as a composite diagnostic, but lacks a consensus implementation. Uniqueness and novelty rely on structure-based fingerprints (e.g., StructureMatcher [Ong et al., 2013]), whose results depend on the reference set, tolerance thresholds, and whether novelty is computed over all samples or only the stable ones. Most works adopt one scheme without benchmarking alternatives or accounting for disorder and symmetry equivalence [Siron et al., 2025], limiting the comparability of reported S.U.N. scores. Additional metrics like **distribution similarity** (e.g., MMD, EMD) and **diversity** are frequently used. However, distribution similarity often rewards memorization rather than novelty—misaligned with discovery goals—while diversity lacks a standardized metric definition or visualizations across studies.

Crucially, most benchmarks focus on **unconditional generation**, where models sample stable crystals broadly from a learned distribution, neglecting the more practically relevant **conditional generation** tasks, where models must satisfy target properties under practical constraints. Even rarer are *limited-budget* settings, where only a fixed number of candidates can be evaluated (i.e. have a ground-truth property), despite being closer to practical applications. **Efficiency** metrics like training time, memory usage, and inference cost are also under-reported, though increasingly important as model sizes grow.

These limitations motivate LEMAT-GENBENCH: a unified, extensible benchmark that proposes a list of standardized metrics and tasks, enabling reproducible model comparison through public tools and a live leaderboard.

3 Benchmark Methodology

We define three benchmark scenarios representing key tasks in crystal generation: unconditional generation, conditional generation, and conditional generation with limited oracle[1] budget. These scenarios correspond to different levels of supervision and real-world applicability, forming the core of LEMAT-GENBENCH.

Unconditional generation trains generative models to produce stable and chemically viable crystalline materials, learning a general-purpose prior over crystalline structures. Once trained, models can be fine-tuned for specific discovery tasks—mirroring workflows in LLMs [Brown et al., 2020, Achiam et al., 2023], diffusion-based image generators [Ramesh et al., 2022], and protein design [Madani et al., 2023]. We propose computing standardized metrics on fixed numbers of generated structures, assessing models' ability to produce realistic, stable, diverse, novel, and unique materials.

Conditional generation evaluates property-conditioned generation with inexpensive oracle calls, without strong resource restrictions (time, compute, oracle queries, labeled data). This benchmark evaluates whether models can produce useful crystals for target applications by optimizing multiple properties (e.g., bandgap, bulk modulus, magnetic density) and accounting for chemical space constraints while maintaining chemically meaningful structures.

Conditional generation with limited budget improves upon the previous scenario by considering expensive oracle calls (e.g., laboratory experiments, high-level quantum calculations) where only small numbers of evaluations are feasible. Here, often starting from pre-trained models, we evaluate sample efficiency in optimizing target properties under oracle budget constraints while maintaining unconditional generation quality (validity, stability, novelty, diversity). These tasks are detailed next.

3.1 Unconditional Generation

To evaluate unconditional generation, a representative number of structures (e.g., 1000) are sampled from the trained model and evaluated using the following metrics, whose formal definitions are provided in Appendix B.

Validity. We define a validity metric, building on Xie et al. [2021], that serves as both a sanity check and a pre-filtering step to exclude malformed or nonphysical structures. These are designed to catch common failure modes and reduce computational overhead. These checks include CIF readability, minimum interatomic distances (>0.5 Å), and density bounds (mass density of 0.01-25 g/cm^3 and atomic density of 10^{-5} and 0.5 atoms/Å^3). We also check for reasonable lattice parameters, a space group compatible with the composition and the lattice parameters, and charge neutrality (via oxidation state plausibility analysis, with a tolerance threshold). Structures failing the validity checks are discarded. All results are aggregated into a single validity score for transparency and comparison. While submitters may pre-filter their structures, the standardized criteria ensure consistency across evaluations.

---

[1] An *oracle* is a function that returns a material property score—whether from an ML model, DFT simulation, or other evaluation method. Oracles vary in computational cost and accuracy.

Stability. Thermodynamic stability is a key proxy for real-world material feasibility, typically assessed via the **energy above the convex hull**, $E_{\text{hull}} = E_{\text{tot}} - E_{\text{hull}}^{\min}$. We consider structures with $E_{\text{hull}} \leq 0$ as stable, and ≤ 0.1 eV/atom as metastable, hoping to harmonize common practices. While formation energy ($E_f = E_{\text{tot}} - \sum_i n_i \mu_i$) has been widely used, it introduces biases—e.g., favoring strongly bonded crystals containing electronegative elements—and is not a reliable proxy for stability. To enable open access scalable evaluation, we compute E_{hull} using an **ensemble of MLIPs** (MACE-MP [Batatia et al., 2023], UMA [Wood et al., 2025], and Orb-v3 [Rhodes et al., 2025]), comparing predicted formation energies to the DFT-based convex hull from LeMat-Bulk [Siron et al., 2025]. While this introduces systematic approximations [Nong et al., 2025], aggregating multiple MLIPs provides robustness and uncertainty estimates. We report the mean and standard deviation of E_{hull} across models. We further evaluate structure quality via a **relaxation check**, computing the root mean square deviation (RMSD) between initial and relaxed atomic positions (averaged over each MLIP). Low RMSD indicates proximity to a local minimum. Note that we report only *direct* S.U.N. scores as participants are encouraged to submit relaxed structures. This combined protocol balances scalability with physical realism and improves comparability across submissions.

Novelty, Uniqueness and Diversity. Exploration capabilities of generative models are assessed through **novelty**, **uniqueness**, and **diversity** metrics. **Novelty** measures the fraction of generated crystals not found in a reference dataset, using structural fingerprints, to identify previously unseen materials. Again, we use LeMat-Bulk [Siron et al., 2025] as reference database, as it represents the most extensive set of known crystal structures, and apply the MatterGen-adapted StructureMatcher [Zeni et al., 2025] as our default, as it handles symmetry, disorder, and known edge cases reliably. To ensure scalability, we restrict comparisons to structures of matching composition, and implement the Short-BAWL fingerprint [Siron et al., 2025] to support scalable comparison across large model output. **Uniqueness** quantifies non-redundancy among generated structures using the same fingerprinting approach. **Diversity** captures the spread of generated samples across crystal sizes, space groups, elemental compositions, and structural descriptors. We compute per-feature Shannon entropy aggregating results into a single diversity score with visualization tools in the LEMAT-GENBENCH codebase. Note that novelty, an approximate metric, should be treated with caution. Structural identity does not always imply functional equivalence, and the boundary between new and known materials is inherently fuzzy. Still, standardizing fingerprint methods, reference datasets, and thresholds is critical for enabling meaningful comparisons and tracking progress. Lastly, note that we also compute Novelty on valid, stable and unique structures, which offers different information compared to doing it on the whole set of generated structures.

(M.)S.U.N. rate. We aggregate core evaluation criteria into a single standardized metric—the S.U.N. rate—which measures the fraction of generated structures that are Stable, Unique, and Novel, computed sequentially from the submitted set in a funnel-like manner. This serves as our primary benchmark for unconditional generation in LEMAT-GENBENCH, with each component precisely defined using fixed thresholds, reference datasets, and fingerprinting methods. To account for synthesizable but metastable materials, we also report the M.S.U.N. rate, relaxing stability to $E_{\text{hull}} \leq 0.1$ eV/atom. This metric sets a practical upper bound on generative performance and provides a consistent, interpretable measure to compare models.

Model efficiency. Beyond output quality, we also report efficiency metrics related to training and inference. Specifically, authors must provide: (i) training compute in FLOPs and time (CPU/GPU days), (ii) inference time to generate 1k relaxed structures on a reference CPU/GPU (e.g. Nvidia A100)[2], and (iii) peak memory usage during inference, together with number of model parameters. Reporting these values will inform model cost and scalability, and support analysis of tradeoffs between performance (e.g., M.S.U.N. rate) and deployment feasibility.

3.2 Conditional Generation

Conditional generation (with inexpensive oracle call) achieves two important goals: it allows downstream users to choose the best-performing model for their task, and gives developers freedom to innovate across all aspects of model development beyond just algorithms. In addition to the uncondi-

---

[2] While inference cost is typically negligible compared to the cost of downstream validation (e.g., DFT or experiments), standardized reporting of these metrics provides important context for practical usability and future integration into closed-loop discovery pipelines.

tional generation metrics, we focus on whether generated crystals exhibit desired properties for the intended application. We detail these additional metrics below.

Property targeting metrics. We evaluate how well models optimize target crystal properties—whether maximizing/minimizing a quantity (e.g., bulk modulus) or targeting a specific range (e.g., bandgap ≈ 2.6 eV). For extremal tasks, we report **top-k values** (for $k = 1, 10, 100$) with mean and standard deviation. For range-targeting tasks, we compute the **success rate**: the fraction of generated structures satisfying the property threshold or interval. While intuitive and useful, these metrics can be gamed through brute-force sampling and filtering. To address this, we report **Conditional (M.)S.U.N. rate**, which evaluates the fraction of stable, unique, and novel structures that satisfy the property constraint. We recommend reporting both top-k scores over the valid (M.)S.U.N. subset and overall success rate to fairly assess generation quality and learning efficiency, ensuring complete evaluation of inverse design tasks.

Multi-objective optimization metrics. Realistic scenarios demand materials that meet multiple property requirements, so models must be evaluated on their ability to optimize these jointly (e.g., bulk modulus, bandgap, HOMO–LUMO, density). To do so, we use **hypervolume indicators** (HVI) as our primary evaluation metric. It quantifies the volume of dominated objective space relative to a reference point, rewarding both high performance and spread of generated samples across conflicting targets. Pareto optimality and multi-objective quality-diversity (MOQD) score [Janmohamed et al., 2024] are also implemented and can be used as diagnosis tests.

Constraint adherence metrics. Besides generating materials with multiple properties of interest, practitioners are interested in focusing on specific compositions (e.g., no rare earth materials), symmetries (e.g. non-centrosymmetric crystals), or other such characteristics. In such constraint-aware scenarios, we additionally report metrics like lattice, space group, and compositional fidelity to assess whether generated materials adhere to target symmetry or elemental design constraints. While these metrics can be applied in unconditional generation settings, conditional generation is application-specific. Therefore, we propose three benchmark tasks to stress-test generative models under realistic, goal-driven discovery settings. Each task is defined in Section 4, with targeted property values and structural constraints.

3.3 Conditional Generation with Limited Budget

In reality, access to labeled data and high-fidelity evaluations are often limited: only a few examples of materials with the desired property or a limited budget to annotate candidates will be available. In such cases, each high-fidelity label (e.g. DFT or experiment) is expensive, making sample-efficient and steerable generation essential. This setting mirrors realistic $R\&D$ workflows, where discovering optimal candidates under strict supervision constraints is a key challenge. This task evaluates how quickly and effectively conditional generative models can identify materials matching a target property, under limited supervision. Inspired by benchmarks in protein and drug design (e.g., TDC Huang et al. [2022], FLIP Dallago et al. [2021]), we adopt a similar philosophy grounded in solid-state materials, tracking **top-k values** and **success rates** under query constraints. We define two complementary setups as follows. In both these cases, the target property will not be revealed to avoid any reward hacking from the community.

- **Offline learning.** A fixed, small dataset (e.g., 1k labeled crystals) is provided to practitioners, with hidden property values computed by a black-box oracle. Generative models can leverage this data freely (e.g., training surrogates, learning property-conditioned latent spaces or designing heuristics) before generating candidates. The evaluation is conducted via the original oracle, using Conditional S.U.N. for top-k or success-based metrics.

- **Online learning** . Practitioners are given access to a black-box oracle API (e.g., MLIP or DFT proxy), which they can query up to a fixed budget (e.g., 1k oracle calls). This setup should reflect realistic optimization workflows, where evaluating each sample incurs a cost and is therefore limited. Practitioners are free to choose what sample they want to label next, incorporating exploration strategies, uncertainty estimation, or adaptive learning to make the most of each query. We track Conditional S.U.N. evolution over query budgets.

4 Towards an Open Benchmark Framework

4.1 Leaderboard implementation

To converge towards a community-wide benchmarking framework, we adopt several concrete steps hopefully accelerating progress in generative crystal modeling. Specifically, we provide a unique **fully open-source implementation of the proposed evaluation metrics**, both for unconditional and conditional evaluation tasks, with and without budget constraints, which can be used and updated by the community as the field evolves. These tools aim to standardize key evaluation components: (i) LeMat-Bulk as the reference dataset for S.U.N, (ii) an ensemble of MLIPs for robust formation energy and convex hull prediction, with uncertainty quantification (iii) validity checks split into hard (pre-filter) and soft (diagnostic) constraints, (iv) standardized diversity scores, (v) resource efficiency reporting, (vi) multi-objective reporting, (viii) pre-relaxed structure submission and RMSD check with a relaxation and (ix) structural equivalence via StructureMatcher and BAWL fingerprints. To further facilitate fair model comparisons, we also introduce a **public leaderboard** on Hugging Face. The submission process is as follows: (i) Authors submit 1k generated crystal structures for unconditional and/or conditional tracks; (ii) Authors may also submit their packaged model, granting a compliance badge (optional); (iii) Reference metrics are computed using our open reference implementation; (iv) Results are displayed with multiple views to support comparison across tasks and generation scenarios. The code is available here `https://github.com/LeMaterial/lemat-genbench`.

For conditional generation (with an inexpensive oracle), we define **three benchmarking tasks**, as indicated in Section 3.2. Remember that for this evaluation suite, we provide the oracle function that measures if generated candidates possess the target property, and that participants are allowed to use this oracle as they wish during training.

1. **Maximize density**. The goal is to generate crystal structures with highest volumetric density. Since density is straightforward to compute from atomic positions and unit cell volume, this task requires no machine-learned potential. It serves as a useful sanity check for structural validity, as models may exploit the objective by tightly packing atoms—potentially violating physical or chemical plausibility.
2. **Bulk modulus**. Target bulk modulus. This benchmark invites models to generate crystals with high bulk modulus ($\geq$ 300GPa), a proxy for incompressibility and super-hard materials like diamond, borides, or carbides [Jin et al., 2023, Kanyanta, 2016]. These are essential for high-pressure and wear-resistant applications. While experimental data is limited, large DFT-labeled datasets (e.g., Materials Project, Matbench) exist, and several MLIPs can estimate this property efficiently. We will evaluate generations using a fixed ML oracle (e.g., MACE), with transparent access to its training data, architecture, and performance.
3. **Stable magnets for sustainable electronics**. This synthetic benchmark is targeted towards rare-earth-free magnetic materials, a task of high technological importance [Vishina et al., 2020, Xia et al., 2022, Kaba et al., 2023]. We would like crystals to exhibit high magnetic density (maximize it), a bandgap between 0.05–0.5 eV, and an Herfindahl–Hirschman Index (HHI) score (an estimate of supply chain risk based on materials availability and cost) lower or equal to 5. This involves multiple property optimization and constraints. The scoring function used to evaluate bandgap and magnetic density will also be released, to provide transparency to practitioners on the leaderboard evaluation.

For conditional generation under limited oracle budget (Section 3.3), we release a list of crystal structures with ground-truth values of a concealed property. Participants operate in a black-box setting that simulates real-world conditions where only a small number of oracle queries (e.g., DFT or experiment) are available. The task setup is deliberately underspecified to discourage reward hacking.

All in all, this evaluation framework and these benchmarking scenarios aim to provide clear, fair, and rigorous standards for evaluating generative models in crystal generation, while remaining flexible to evolving research needs and application contexts. It should enable more meaningful and fine-grained comparisons between crystal generative approaches, highlight promising research directions, and ultimately help bridge the gap between computational prediction and experimental realization. This infrastructure is particularly crucial as the field moves toward closed-loop discovery pipelines that integrate computational prediction, experimental validation, and synthesis planning. More details can be found in the LEMAT-GENBENCH codebase and on the Hugging Face leaderboard. Finally, let's emphasize that, not unlike the rest of the field, this evaluation framework and benchmark has many limitations and areas for future development; nevertheless, we seek to provide a first step towards accelerating progress in materials generation through better model evaluation practices.

4.2 Benchmarking workflow

To ensure consistency, scalability, and physical soundness in evaluation, our benchmark pipeline consists of two main phases: (i) mandatory validity filtering and (ii) metric-specific preprocessing and evaluation. The full implementation is in Appendix C and the public codebase. For now, we focus on unconditional generation. Conditional benchmarks will be made public upon paper acceptance.

Phase 1: Validity Filtering. All submitted structures undergo a standardized validity check before any downstream metric is computed. This includes CIF readability, minimum interatomic distance thresholds, physical density bounds, charge neutrality and symmetry consistency. Structures that fail these constraints are excluded from further analysis, ensuring that downstream metrics (e.g., energy above hull or diversity) are computed only on physically plausible samples. This filtering step reduces computational overhead and serves as a common sanity layer across models.

Phase 2: Metric-Specific Preprocessing. Depending on the evaluation metrics selected, the system dynamically applies a set of modular pre-processors. For example:

- Stability pre-processors use an ensemble of MLIPs (ORB v3, MACE-MP, UMA) to relax structures, compute formation energies, and evaluate energy above hull.
- Fingerprint-based metrics (e.g., novelty, uniqueness, S.U.N.) use either hash-based fingerprints (short-BAWL) or direct pairwise comparison via `StructureMatcher`. The `FingerprintPreprocessor` is applied only for fingerprint-based methods, and skipped when using `StructureMatcher`.
- Distribution pre-processor (e.g., MMD, JS divergence) computes structural and compositional statistics to compare against the reference LeMat-Bulk dataset.

Each preprocessor attaches computed features to the structure objects, enabling smooth integration with subsequent benchmarks. Benchmarks are executed independently with optimized memory management. Outputs are stored in a structured JSON format with full traceability and diagnostics.

By decoupling filtering, preprocessing, and evaluation, the benchmark is modular, extensible, and robust to different modeling paradigms.

4.3 Benchmarking Results and Discussion

We evaluate ten generative models for crystalline materials using the LEMAT-GENBENCH benchmark. These include: ADiT [Joshi et al., 2025], Crystalformer [Cao et al., 2024], DiffCSP [Jiao et al., 2023], DiffCSP++ [Jiao et al., 2024], LLaMat2 [Mishra et al., 2024], MatterGen [Zeni et al., 2025], PLaID++ [Xu et al., 2025b], SymmCD [Levy et al., 2025b], WyFormer [Kazeev et al., 2025], and WyFormer-DFT [Kazeev et al., 2025]. All models were trained on the MP-20 dataset; most rely on diffusion or autoregressive backbones, while PLaID++ additionally leverages reinforcement learning. For each model, we evaluate 1,000 structures either obtained from the authors or sourced from public repositories [Kazeev et al., 2025]. A detailed breakdown of data sources is provided in Table 4.

All evaluations are performed on the submitted structures as is, *without re-relaxation*. This choice simplifies benchmarking and ensures strict reproducibility but may underestimate metrics such as S.U.N. and M.S.U.N., which are sensitive to structural relaxation. We encourage future submissions to include relaxed structures to better capture thermodynamic viability. Post-relaxation metrics are currently available for indicative analysis and will be formally integrated in upcoming benchmark releases.

Importantly, we use LeMat-Bulk as the reference dataset rather than MP-20, which serves as the training set for all evaluated models (Table 4). This distinction reflects our focus on downstream recovery performance rather than model-specific generative capabilities.

Model	Valid ↑	Unique ↑	Novel ↑	Energy-based ↓			Stable ↑	SUN ↑	MSUN ↑
				E_f	E_hull	Relax-RMSD			
ADiT	905	897	277	-0.66 ± 0.91	0.41 ± 0.54	0.39 ± 0.41	$\underline{25}$	2	5
Crystalformer	699	694	318	-0.17 ± 1.48	0.74 ± 1.35	0.66 ± 1.07	16	6	14
DiffCSP	961	$\underline{958}$	$\underline{666}$	$\underline{-0.67 \pm 0.83}$	0.32 ± 0.29	0.60 ± 0.64	19	**12**	32
DiffCSP++	957	956	652	-0.54 ± 0.86	0.47 ± 0.46	0.73 ± 0.83	23	$\underline{11}$	22
LLaMat2	860	848	331	-0.47 ± 1.06	0.47 ± 0.79	0.52 ± 0.66	22	6	16
SymmCD	723	722	476	-0.02 ± 1.18	0.92 ± 1.05	0.89 ± 1.06	13	5	11
MatterGen	$\underline{962}$	**961**	**700**	$\mathbf{-0.71 \pm 0.79}$	$\underline{0.24 \pm 0.24}$	$\underline{0.37 \pm 0.42}$	$\underline{25}$	$\underline{11}$	**66**
PLaID++	**985**	873	241	-0.50 ± 0.46	$\mathbf{0.13 \pm 0.22}$	$\mathbf{0.12 \pm 0.24}$	**29**	3	$\underline{51}$
WyFormer	938	938	650	-0.47 ± 0.93	0.52 ± 0.50	0.75 ± 0.79	18	7	7
WyFormer-DFT	951	946	662	-0.65 ± 0.88	0.35 ± 0.36	0.40 ± 0.61	17	10	32

Table 1: Core unconditional generation benchmark metrics for 10 generative crystal models evaluated on 1,000 generated structures each. Arrows indicate optimization direction. Bold indicates best performance, underlined indicates second-best. The energy-based metrics are computed over the full generated sets, and the relatively high standard deviations reflect the stochasticity in the structural generation process rather than measurement error. Models below the separator (MatterGen, PLaID++, WyFormer, and WyFormer-DFT) incorporate structural relaxation within their generation or evaluation pipelines. All reported metrics are computed directly on the raw, submitted structures as provided by the models.

Key takeaways. Table 1 and Table 2 reveal substantial diversity in model performance across evaluation axes, with no single model dominating all metrics. This heterogeneity underscores the critical need for multi-faceted benchmarking in generative materials design: different downstream applications demand different priorities.

MatterGen achieves the highest **M.S.U.N.** rate (66), positioning it as the strongest candidate for discovering metastable yet unique and novel materials. **DiffCSP** leads in **S.U.N.** (12), demonstrating superior capability in generating stable, unique, and novel structures. Notably, both **DiffCSP** and **DiffCSP++** achieve these strong S.U.N. scores *without incorporating structural relaxation* into their generation pipelines, highlighting the intrinsic physical consistency of their diffusion-based priors. In contrast, **PLaID++** attains the highest validity (985) and stability (29 stable structures), but exhibits more moderate novelty and S.U.N. rates, reflecting a narrower generative distribution. This behavior aligns with its **reinforcement learning** objective, which explicitly penalizes unstable outputs through Direct Preference Optimization.

Energy-based metrics reveal contrasting strengths. MatterGen and DiffCSP achieve the lowest formation energies (E_f of -0.71 and -0.67 eV respectively), while PLaID++ leads on energy above hull (0.13 eV), a result of its DPO training that directly optimizes for thermodynamic stability through MLIP-based reward signals and integrated relaxation steps in its generation pipeline. Relaxation RMSD, a proxy for how close structures are to local energy minima, is lowest for PLaID++ (0.12 Å) and MatterGen (0.37 Å), indicating these models generate structures near equilibrium geometries. PLaID++'s advantage here is amplified by its use of relaxation as part of the generation workflow. This is to be expected for a sample of pre-relaxed structures, and it is therefore quite surprising that the relaxation RMSE is so comparatively high for WyFormer, with 0.75 being higher than 5/6 of the un-relaxed samples.

Distribution similarity metrics displayed in Table 2 (JS divergence, MMD, FID) often trade off with novelty and diversity. SymmCD achieves the best JS divergence (0.214) through its symmetry-preserving constraints, but exhibits lower S.U.N./M.S.U.N. rates, suggesting potential overfitting to training distribution patterns. In contrast, diffusion-based models like DiffCSP and MatterGen show higher distributional divergence but achieve superior novelty and metastability scores, demonstrating that broader chemical space exploration necessarily deviates from training distributions. DiffCSP++ and WyFormer-DFT balance this trade-off effectively, with WyFormer-DFT showing strong uniqueness, competitive energy metrics, and high novelty, validating the utility of DFT-relaxed structure generation as supervision signal.

WyFormer variants demonstrate the strongest diversity metrics, with WyFormer achieving the best SiteDiv (23.083) and ElemDiv (0.729), while WyFormer-DFT leads in space group diversity (SGDiv: 0.417). These Wyckoff-based transformer models effectively explore diverse symmetry groups and compositional spaces. SymmCD and Crystalformer also show strong performance across multiple diversity axes, though with different strengths: Crystalformer excels in size diversity (0.313) while

SymmCD balances elemental and space group coverage. Models exhibiting higher diversity generally show elevated JS/FID scores, confirming that extensive chemical exploration comes at the cost of training distribution fidelity.

These results illuminate how architectural choices drive performance trade-offs. LLM-based approaches (PLaID++) excel at validity and stability through preference optimization but show constrained exploration. Diffusion models (DiffCSP, MatterGen, DiffCSP++) balance stability with higher novelty through their probabilistic sampling mechanisms, with symmetry-aware variants (DiffCSP++, SymmCD) showing improved physical validity. Wyckoff-based representations (PLaID++, WyFormer variants) consistently enhance compositional validity and space group diversity by encoding crystallographic constraints directly into the generation process. These findings emphasize that model selection must align with specific research objectives, whether prioritizing exploration (MatterGen, DiffCSP), safety and stability (PLaID++), or diversity (WyFormer variants). The standardized evaluation framework provided by LeMat-GenBench enables researchers to make these informed decisions and identify opportunities for architectural improvements.

Model	**Distribution**			**Diversity**				**HHI**	
	JS ↓	**MMD ↓**	**FID ↓**	**ElemDiv ↑**	**SGDiv ↑**	**SizeDiv ↑**	**SiteDiv ↑**	**Prod ↓**	**Res ↓**
ADiT	0.517	**0.002**	1.797	0.703	0.022	0.266	14.585	3.473	2.731
Crystalformer	0.255	0.003	2.313	0.712	0.352	**0.313**	18.586	3.784	2.821
DiffCSP	0.457	0.008	1.743	0.712	0.109	0.272	14.806	**3.327**	**2.605**
DiffCSP++	0.221	0.004	1.517	0.712	**0.417**	0.299	21.346	3.533	2.759
LLaMat2	0.320	0.003	**1.377**	0.712	0.191	0.267	9.755	3.975	3.013
SymmCD	**0.214**	0.006	2.345	0.720	0.404	0.311	19.773	3.496	2.668
MatterGen	0.440	0.006	1.741	0.644	0.130	0.268	12.948	3.484	2.682
PLaID++	0.437	0.034	2.875	0.669	0.217	0.238	6.144	5.238	3.410
WyFormer	0.224	0.008	1.400	**0.729**	0.400	0.303	**23.083**	3.533	2.724
WyFormer-DFT	0.265	0.010	2.028	0.720	**0.417**	0.298	22.223	3.506	2.744

Table 2: Distribution similarity, diversity, and supply-chain risk metrics for generative crystal models. Lower values are better for distribution metrics — JS (Jensen–Shannon divergence), MMD (Maximum Mean Discrepancy), and FID (Fréchet Inception Distance) — and for HHI (Herfindahl–Hirschman Index) metrics. Higher values indicate greater diversity across ElemDiv (elemental diversity), SGDiv (space-group diversity), SizeDiv (crystal size diversity), and SiteDiv (atomic site diversity). "Prod" and "Res" denote production- and reserve-side HHI components, respectively. Models below the separator (MatterGen, PLaID++, WyFormer, and WyFormer-DFT) incorporate structural relaxation in their generation or evaluation pipelines.

5 Conclusions and Outlook

Generative models for crystalline materials are rapidly reshaping the landscape of computational discovery, but their evaluation remains inconsistent and fragmented. LeMat-GenBench addresses this gap by providing a unified, extensible framework for benchmarking generative models of inorganic crystals. It standardizes a core set of metrics centered on stability, uniqueness, and novelty (S.U.N.), together with validity, diversity, and efficiency. These metrics are complemented by a suite of evaluation tasks beyond unconditional generation, including conditional and limited-budget settings, bringing assessment practices closer to real-world discovery scenarios.

We release an open-source evaluation toolkit and public leaderboard enabling model comparisons under consistent protocols. Baseline results from 10 recent models demonstrate LeMat-GenBench's value in diagnosing performance trade-offs and revealing model patterns. Our results show no single model dominates all metrics. By aligning model evaluation with realistic constraints and application needs, LeMat-GenBench aims to guide development of more capable, reliable, and scientifically useful generative models. We see this as a first step toward closing the loop between computational generation and experimental validation.

Limitations and Future Directions. This release focuses on unconditional generation evaluation, with conditional generation benchmarks and expanded model implementations planned for future updates. Key challenges remain. Data quality is a persistent bottleneck: widely used datasets often lack compositional diversity, structural metadata, or negative examples necessary for robust training. Property-conditioned generation is further hindered by unreliable surrogate models and inconsistent conditioning protocols. Most generative models assume idealized, defect-free crystals, overlooking critical phenomena like disorder, doping, and non-stoichiometry that shape real-world functionality. Moreover, stability assessment relies on ensembles of MLIPs, which—despite averaging—can deviate systematically from DFT, especially near stability thresholds. Finally, although we assess thermodynamic plausibility, our framework does not yet capture kinetic barriers, synthesis feasibility, or real-world constraints. Bridging these gaps—especially toward synthesis-aware and experimentally grounded pipelines—will require tighter integration between data, modeling, and validation across disciplines. Environmental and sustainability considerations are discussed in Appendix D.

6 Acknowledgments

We acknowledge the fruitful discussions and insights brought by Alfonso Amayuelas, Anuroop Sriram, Benjamin Kurt Miller, Edvin Fako, Hashim Piracha, Kishalay Das, Victor Schmidt, Zekun Danny Ren, Aron Walsh and Tian Xie.

References

Jonathan Schmidt, Mário RG Marques, Silvana Botti, and Miguel AL Marques. Recent advances and applications of machine learning in solid-state materials science. *npj computational materials*, 5 (1):83, 2019.

Walter Kohn, Axel D Becke, and Robert G Parr. Density functional theory of electronic structure. *The journal of physical chemistry*, 100(31):12974–12980, 1996.

David S. Sholl and Janice A. Steckel. *Density Functional Theory: A Practical Introduction*. Wiley, March 2009. ISBN 9780470447710. doi:10.1002/9780470447710. URL `http://dx.doi.org/10.1002/9780470447710`.

Volker L Deringer, Miguel A Caro, and Gábor Csányi. Machine learning interatomic potentials as emerging tools for materials science. *Advanced Materials*, 31(46):1902765, 2019.

Oliver T Unke, Stefan Chmiela, Huziel E Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T Schutt, Alexandre Tkatchenko, and Klaus-Robert Muüller. Machine learning force fields. *Chemical Reviews*, 121(16):10142–10186, 2021.

Alexandre Duval, Simon V Mathis, Chaitanya K Joshi, Victor Schmidt, Santiago Miret, Fragkiskos D Malliaros, Taco Cohen, Pietro Liò, Yoshua Bengio, and Michael Bronstein. A hitchhiker's guide to geometric gnns for 3d atomic systems. *arXiv preprint arXiv:2312.07511*, 2023a.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, et al. Mattergen: a generative model for inorganic materials design. *arXiv preprint arXiv:2312.03687*, 2023.

Mila AI4Science, Alex Hernandez-Garcia, Alexandre Duval, Alexandra Volokhova, Yoshua Bengio, Divya Sharma, Pierre Luc Carrier, Michał Koziarski, and Victor Schmidt. Crystal-GFN: sampling crystals with desirable properties and constraints. *AI for Accelerated Materials Design Workshop (NeurIPS)*, 2023.

Daniel Levy, Siba Smarak Panigrahi, Sékou-Oumar Kaba, Qiang Zhu, Kin Long Kelvin Lee, Mikhail Galkin, Santiago Miret, and Siamak Ravanbakhsh. Symmcd: Symmetry-preserving crystal generation with diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Nikita Kazeev, Wei Nong, Ignat Romanov, Ruiming Zhu, Andrey E Ustyuzhanin, Shuya Yamazaki, and Kedar Hippalgaonkar. Wyckoff transformer: Generation of symmetric crystals. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=eFHfRQRjJo`.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013. URL `https://arxiv.org/abs/1312.6114`.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=PxTIG12RRHS`.

Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J Hu, Mo Tiwari, and Emmanuel Bengio. GFlowNet Foundations. *Journal of Machine Learning Research (JMLR)*, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Samuel M. Blau, Vlad Cărare, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Fabio Falcioni, Edvin Fako, Andrea C. Ferrari, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Christian Holm, Jad Jaafar, Stephan Hofmann, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Viktor Svahn, Christopher Sutton, Thomas D. Swinburne, Jules Tilly, Cas van der Oord, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry. *arXiv preprint arXiv: 2401.00096*, 2023.

Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):138, 2020.

Juhwan Noh, Jaehoon Kim, Helge S. Stein, Benjamin Sanchez-Lengeling, John M. Gregoire, Alan Aspuru-Guzik, and Yousung Jung. Inverse design of solid-state materials via a continuous representation. *Matter*, 1(5):1370–1384, 2019. ISSN 2590-2385. doi:10.1016/j.matt.2019.08.017.

Jordan Hoffmann, Louis Maestrati, Yoshihide Sawada, Jian Tang, Jean Michel Sellier, and Yoshua Bengio. Data-driven approach to encoding and decoding 3-d crystal structures, 2019. URL `http://arxiv.org/abs/1909.00949`.

Callum J. Court, Batuhan Yildirim, Apoorv Jain, and Jacqueline M. Cole. 3-d inorganic crystal structure generation and property prediction via representation learning. *Journal of Chemical Information and Modeling*, 60(10):4518–4535, 2020. ISSN 1549-9596, 1549-960X. doi:10.1021/acs.jcim.0c00464. URL `https://pubs.acs.org/doi/10.1021/acs.jcim.0c00464`.

Yong Zhao, Edirisuriya M. Dilanga Siriwardane, Zhenyao Wu, Nihang Fu, Mohammed Al-Fahdi, Ming Hu, and Jianjun Hu. Physics guided deep learning for generative design of crystal materials with symmetry constraints. *npj Computational Materials*, 9(1):38, 2023. ISSN 2057-3960. doi:10.1038/s41524-023-00987-9. URL `https://doi.org/10.1038/s41524-023-00987-9`.

Sungwon Kim, Juhwan Noh, Geun Ho Gu, Alan Aspuru-Guzik, and Yousung Jung. Generative adversarial networks for crystal structure prediction. *ACS Central Science*, 6(8):1412–1420, 2020. doi:10.1021/acscentsci.0c00426. URL `https://doi.org/10.1021/acscentsci.0c00426`. PMID: 32875082.

Yong Zhao, Mohammed Al-Fahdi, Ming Hu, Edirisuriya M. D. Siriwardane, Yuqi Song, Alireza Nasiri, and Jianjun Hu. High-throughput discovery of novel cubic crystal materials using deep generative neural networks. *Advanced Science*, 8(20):2100566, 2021. doi:https://doi.org/10.1002/advs.202100566. URL `https://advanced.onlinelibrary.wiley.com/doi/abs/10.1002/advs.202100566`.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*, 2021.

Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. *Advances in Neural Information Processing Systems*, 36:17464–17497, 2023.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, et al. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, 2025.

Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation. *arXiv preprint arXiv:2402.03992*, 2024.

Rees Chang, Angela Pak, Alex Guerra, Ni Zhan, Nick Richardson, Elif Ertekin, and Ryan P Adams. Space group equivariant crystal diffusion. *arXiv preprint arXiv:2505.10994*, 2025.

Kishalay Das, Subhojyoti Khastagir, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. Periodic materials generation using text-guided joint diffusion model. In *The Thirteenth International Conference on Learning Representations*, 2025.

Hyunsoo Park, Anthony Onwuli, and Aron Walsh. Exploration of crystal chemical space using text-guided generative artificial intelligence. *Nature Communications*, 16(1):4379, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=PqvMRDCJT9t`.

Benjamin Miller, Ricky Chen, Anuroop Sriram, and Brandon Wood. Flowmm: Generating materials with riemannian flow matching. In *International Conference on Machine Learning (ICML)*. ICML, jun 2024.

Anuroop Sriram, Benjamin Miller, Ricky Chen, and Brandon Wood. Flowllm: Flow matching for material generation with large language models as base distributions. In *Neural Information Processing Systems (NeurIPS)*. NeurIPS, jan 2024.

Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C Lawrence Zitnick, and Zachary Ulissi. Fine-tuned language models generate stable inorganic materials as text. *arXiv preprint arXiv:2402.04379*, 2024.

Chaitanya K Joshi, Xiang Fu, Yi-Lun Liao, Vahe Gharakhanyan, Benjamin Kurt Miller, Anuroop Sriram, and Zachary W Ulissi. All-atom diffusion transformers: Unified generative modelling of molecules and materials. In *International Conference on Learning Representations*, 2025.

Omri Puny, Yaron Lipman, and Benjamin Kurt Miller. Space group conditional flow matching. *arXiv preprint arXiv:2509.23822*, 2025.

Elena Zamaraeva, Christopher M Collins, Dmytro Antypov, Vladimir V Gusev, Rahul Savani, Matthew S Dyer, George R Darling, Igor Potapov, Matthew J Rosseinsky, and Paul G Spirakis. Reinforcement learning in crystal structure prediction. *Digital Discovery*, 2(6):1831–1840, 2023.

Prashant Govindarajan, Santiago Miret, Jarrid Rector-Brooks, Mariano Phielipp, Janarthanan Rajendran, and Sarath Chandar. Learning conditional policies for crystal design using offline reinforcement learning. *Digital Discovery*, 3(4):769–785, 2024.

Flaviu Cipcigan, Jonathan Booth, Rodrigo Neumann Barros Ferreira, Carine Ribeiro Dos Santos, and Mathias Steiner. Discovery of novel reticular materials for carbon dioxide capture using GFlowNets. *Digital Discovery*, 2024.

Lena Podina, Christina Humer, Alexandre Duval, Victor Schmidt, Ali Ramlaoui, Shahana Chatterjee, Yoshua Bengio, Alex Hernandez-Garcia, David Rolnick, and Félix Therrien. Catalyst gflownet for electrocatalyst design: A hydrogen evolution reaction case study. *arXiv preprint arXiv:2510.02142*, 2025.

Andy Xu, Rohan Desai, Larry Wang, Gabriel Hope, and Ethan T Ritz. Plaid: Preference aligned language model for targeted inorganic materials design. In *AI for Accelerated Materials Design-ICLR*, 2025a.

Frederik Lizak Johansen, Ulrik Friis-Jensen, Erik Bjørnager Dam, Kirsten Marie Ørnsbjerg Jensen, Rocío Mercado, and Raghavendra Selvan. decifer: Crystal structure prediction from powder diffraction data using autoregressive language models. *arXiv preprint arXiv:2502.02189*, 2025.

Viggo Moro, Charlotte Loh, Rumen Dangovski, Ali Ghorashi, Andrew Ma, Zhuo Chen, Samuel Kim, Peter Y Lu, Thomas Christensen, and Marin Soljačić. Multimodal foundation models for material property prediction and discovery. *Newton*, 2025.

Xiang Fu, Zhenghao Wu, Wujie Wang, Tian Xie, Sinan Keten, Rafael Gomez-Bombarelli, and Tommi Jaakkola. Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *arXiv preprint arXiv:2210.07237*, 2022.

Wei Nong, Ruiming Zhu, Zekun Ren, Martin Hoffmann Petersen, Shuya Yamazaki, Nikita Kazeev, Andrey Ustyuzhanin, Gang Wu, Shuo-Wang Yang, and Kedar Hippalgaonkar. Energy underprediction from symmetry in machine-learning interatomic potentials. *arXiv preprint arXiv:2507.15190*, 2025.

Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L Chevrier, Kristin A Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013.

Martin Siron, Inel Djafar, Etienne du Fayet, Amandine Rossello, Ali Ramlaoui, and Alexandre Duval. Lemat-bulk: aggregating, and de-duplicating quantum chemistry materials databases. In *AI for Accelerated Materials Design-ICLR 2025*, 2025.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature biotechnology*, 41(8): 1099–1106, 2023.

Brandon M Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John R Kitchin, Daniel S Levine, et al. Uma: A family of universal models for atoms. *arXiv preprint arXiv:2506.23971*, 2025.

Benjamin Rhodes, Sander Vandenhaute, Vaidotas Šimkus, James Gin, Jonathan Godwin, Tim Duignan, and Mark Neumann. Orb-v3: atomistic simulation at scale. *arXiv preprint arXiv:2504.06231*, 2025.

Hannah Janmohamed, Marta Wolinska, Shikha Surana, Thomas Pierrot, Aron Walsh, and Antoine Cully. Multi-objective quality-diversity for crystal structure prediction. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1273–1281, 2024.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Artificial intelligence foundation for therapeutic science. *Nature Chemical Biology*, 18(10):1033–1036, 2022.

Christian Dallago, Jody Mou, Kadina E Johnston, Bruce J Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K Yang. FLIP: Benchmark tasks in fitness landscape inference for proteins. *bioRxiv*, pages 2021–11, 2021.

Ruihua Jin, Xiaoang Yuan, and Enlai Gao. Atomic stiffness for bulk modulus prediction and high-throughput screening of ultraincompressible crystals. *Nature Communications*, 14(1):4258, 2023.

Valentine Kanyanta. Hard, superhard and ultrahard materials: An overview. *Microstructure-Property Correlations for Hard, Superhard, and Ultrahard Materials*, pages 1–23, 2016.

Alena Vishina, Olga Yu Vekilova, Torbjörn Björkman, Anders Bergman, Heike C Herper, and Olle Eriksson. High-throughput and data-mining approach to predict new rare-earth free permanent magnets. *Physical Review B*, 101(9):094407, 2020.

Weiyi Xia, Masahiro Sakurai, Balamurugan Balasubramanian, Timothy Liao, Renhai Wang, Chao Zhang, Huaijun Sun, Kai-Ming Ho, James R Chelikowsky, David J Sellmyer, et al. Accelerating the discovery of novel magnetic materials using machine learning–guided adaptive feedback. *Proceedings of the National Academy of Sciences*, 119(47):e2204485119, 2022.

Sékou-Oumar Kaba, Benjamin Groleau-Paré, Marc-Antoine Gauthier, A-MS Tremblay, Simon Verret, and Chloé Gauvin-Ndiaye. Prediction of large magnetic moment materials with graph neural networks and random forests. *Physical Review Materials*, 7(4):044407, 2023.

Zhendong Cao, Xiaoshan Luo, Jian Lv, and Lei Wang. Space group informed transformer for crystalline materials generation. *arXiv preprint arXiv:2403.15734*, 2024.

Vaibhav Mishra, Somaditya Singh, Dhruv Ahlawat, Mohd Zaki, Vaibhav Bihani, Hargun Singh Grover, Biswajit Mishra, Santiago Miret, N M Anoop Krishnan, et al. Foundational large language models for materials research. *arXiv preprint arXiv:2412.09560*, 2024.

Andy Xu, Rohan Desai, Larry Wang, Gabriel Hope, and Ethan Ritz. Plaid++: A preference aligned language model for targeted inorganic materials design. *arXiv preprint arXiv:2509.07150*, 2025b.

Daniel Levy, Siba Smarak Panigrahi, Sékou-Oumar Kaba, Qiang Zhu, Kin Long Kelvin Lee, Mikhail Galkin, Santiago Miret, and Siamak Ravanbakhsh. Symmcd: Symmetry-preserving crystal generation with diffusion models. *arXiv preprint arXiv:2502.03638*, 2025b.

Dejan Zagorac, H Müller, S Ruehl, J Zagorac, and Silke Rehme. Recent developments in the inorganic crystal structure database: theoretical crystal structure data and related features. *Applied Crystallography*, 52(5):918–925, 2019.

Martin Siron, Inel Djafar, Lucile Ritchie, Etienne Du-Fayet, Amandine Rossello, Ali Ramlaoui, Leandro von Werra, Thomas Wolf, and Alexandre Duval. Lemat-bulk dataset, 2024. URL `https://huggingface.co/datasets/LeMaterial/LeMat-Bulk`.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

Bent Fuglede and Flemming Topsoe. Jensen-shannon divergence and hilbert space embedding. In *International symposium onInformation theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE, 2004.

Ilya O Tolstikhin, Bharath K Sriperumbudur, and Bernhard Schölkopf. Minimax estimation of maximum mean discrepancy with radial kernels. *Advances in Neural Information Processing Systems*, 29, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Gunter Klambauer. Fréchet ChemNet distance: a metric for generative models for molecules in drug discovery. *Journal of Chemical Information and Modeling*, 58(9):1736–1741, 2018.

Wenhao Gao, Tianfan Fu, Jimeng Sun, and Connor Coley. Sample efficiency matters: a benchmark for practical molecular optimization. *Advances in neural information processing systems*, 35: 21342–21357, 2022.

Alexandre Agm Duval, Victor Schmidt, Alex Hernández-Garcıa, Santiago Miret, Fragkiskos D Malliaros, Yoshua Bengio, and David Rolnick. Faenet: Frame averaging equivariant gnn for materials modeling. In *International Conference on Machine Learning*, pages 9013–9033. PMLR, 2023b.

Alexandre Duval, Victor Schmidt, Santiago Miret, Yoshua Bengio, Alex Hernández-García, and David Rolnick. Phast: Physics-aware, scalable, and task-specific gnns for accelerated catalyst design. *Journal of Machine Learning Research*, 25(106):1–26, 2024.

Seonghwan Seo, Minsu Kim, Tony Shen, Martin Ester, Jinkyoo Park, Sungsoo Ahn, and Woo Youn Kim. Generative flows on synthetic pathway for drug design. *arXiv preprint arXiv:2410.04542*, 2024.

Jeff Guo and Philippe Schwaller. Directly optimizing for synthesizability in generative molecular design using retrosynthesis models. *Chemical Science*, 2025.

Wenhao Gao, Shitong Luo, and Connor W Coley. Generative artificial intelligence for navigating synthesizable chemical space. *arXiv preprint arXiv:2410.03494*, 2024.

Daniel Widdowson and Vitaliy Kurlin. Pointwise distance distributions for detecting near-duplicates in large materials databases. *arXiv preprint arXiv: 2108.04798*, 2021.

Supplementary Material

Box 1: Key Terms in Generative Modeling

Generative Model: A machine learning model that learns a data distribution $p(\mathbf{x})$ (or a conditional distribution $p(\mathbf{x}|\mathbf{z})$ or $p(\mathbf{x}|\mathbf{c})$) and can generate new samples $\mathbf{x}' \sim p(\mathbf{x})$ that resemble the training data.

Latent Space: A lower-dimensional representation space $\mathbf{z} \in \mathbb{R}^d$ learned by models such as VAEs or GANs, where semantic attributes of the data are often encoded.

Prior Distribution: A predefined distribution (e.g., Gaussian) over the latent variables, typically denoted as $p(\mathbf{z})$, from which samples are drawn during generation.

Decoder / Generator: A neural network (often denoted $G(\mathbf{z})$) that maps latent codes $\mathbf{z}$ to data samples $\mathbf{x}$.

Reconstruction Loss: A metric used in training autoencoders and VAEs that measures how well the generated sample $\hat{\mathbf{x}}$ matches the original input $\mathbf{x}$:

$$\mathcal{L}_{\text{recon}} = \|\hat{\mathbf{x}} - \mathbf{x}\|^2 \quad \text{or} \quad -\log p(\mathbf{x}|\mathbf{z}).$$

KL Divergence: A measure of how much one probability distribution differs from another. Commonly used in VAEs to regularize the encoder:

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}(q(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})).$$

Mode Collapse: A failure mode in GANs where the generator produces samples with limited diversity, collapsing to a few modes of the data distribution.

Conditional Generation: Generation of samples $\mathbf{x}$ based on specified properties or constraints $\mathbf{c}$, e.g., $p(\mathbf{x}|\mathbf{c})$, enabling property-guided design.

Inverse Design: The process of searching the input space (e.g., structure, composition) that maps to a desired target property, often using a generative model or an optimization loop in latent space.

Diffusion Models: A class of generative models that learn to reverse a stochastic diffusion process. Data $\mathbf{x}_0$ is gradually perturbed into noise via:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_0, (1-\alpha_t)I).$$

and a neural network is trained to denoise $\mathbf{x}_t$ to recover $\mathbf{x}_0$ through a learned reverse process $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$.

Score-Based Models: Closely related to diffusion models, they learn the score function $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ and use Langevin dynamics or ODE solvers to sample from the data distribution.

$$\log p(\mathbf{x}) = \log p(\mathbf{z}) + \sum_{k=1}^{K} \log \left| \det \frac{\partial f_k^{-1}}{\partial \mathbf{x}} \right|.$$

Flow Matching: A recent generative approach that avoids training score functions or simulating diffusion. It directly learns a vector field $\mathbf{v}_\theta(\mathbf{x}, t)$ that maps noise to data through an ODE:

$$\frac{d\mathbf{x}}{dt} = \mathbf{v}_\theta(\mathbf{x}, t).$$

This method can be trained via supervised learning on synthetic trajectories or velocity fields between the base and target distributions.

Box 2: Key Terms in Crystallography & Materials Science

Crystal Lattice: A crystal structure is periodic in three dimensions. This periodicity is described by the lattice, which is defined as

$$\mathbf{L} = \{l_1\mathbf{a}_1 + l_2\mathbf{a}_2 + l_1\mathbf{a}_3 | l_1, l_2, l_3 \in \mathbb{Z}\},$$

where a_1, a_2, a_3 are basis vectors of $\mathbb{R}^3$.

Unit Cell: A unit cell is the smallest unit that can be translated to define the whole lattice. In three dimensions, it is always a parallelepiped.

Lattice Parameters: A lattice is typically defined in two ways: either as a set of three basis vectors, or as a set of lattice parameters $(a, b, c, \alpha, \beta, \gamma)$, where a, b, c are the lengths of edges of the unit cell, and α, β, γ are the angles between them.

Symmetry: An object's symmetry is given by the set of geometric transformations that map the object onto itself, leaving it invariant.

Space Group: Crystals can be classified by their symmetries. They possess the translational symmetry of their crystal lattices, and they may also have the point group symmetries of rotations and reflections within a unit cell. The combination of translational and point group symmetries can yield more transformations that a crystal can be symmetric to, including screw and glide symmetries. The full set of symmetric transformations that leave a crystal invariant defines the space group of the crystal. In three dimensions, there are 230 types of space groups.

Wyckoff Position: Applying symmetry operations to a crystal may leave some atoms unaffected: for example, a rotation about an axis leaves atoms on the axis in the same position. The set of symmetry operations that do not move a position is that position's site symmetry. A Wyckoff position is a set of positions that all have the same site symmetries, or conjugate site symmetries. For example, all points along a mirror plane may belong to the same Wyckoff position, while a point at the origin of a unit cell may have its own Wyckoff position. Every point in a crystal can be assigned a Wyckoff position.

Formation Energy: The formation energy of a crystal is the difference in energy between the crystal and its constituent elements.

Energy above Convex Hull: The convex hull gives linear combinations of known phases that represent the lowest-energy mixtures of materials; if a material has an energy above the hull ($E_{\text{hull}} > 0$), it is energetically favorable for it to decompose into a combination of stable phases and is therefore thermodynamically unstable. For example, the convex hull of table salt, NaCl, also includes pure stable Na, pure Cl, as well as $NaCl_3$. However, Na_2Cl has a higher formation energy than the combination of NaCl and pure Na, so it is unstable.

Metastable: Even if a crystal is not in its lowest possible energy state, it may still be metastable, meaning that a potential energy barrier prevents it from easily transitioning to a lower-energy state. A crystal having a low energy above the convex hull while also being at an energy minimum may indicate that it is metastable. Metastable materials are still important: for example, diamond is metastable, but does not readily convert to a lower energy state under normal conditions.

Band Gap: The band gap is the difference in energy between the valence band and the conduction band in a solid.

CIF: Crystallographic Information File, a string-based encoding of a crystal that includes information such as atom positions, unit cell parameters, and chemical elements.

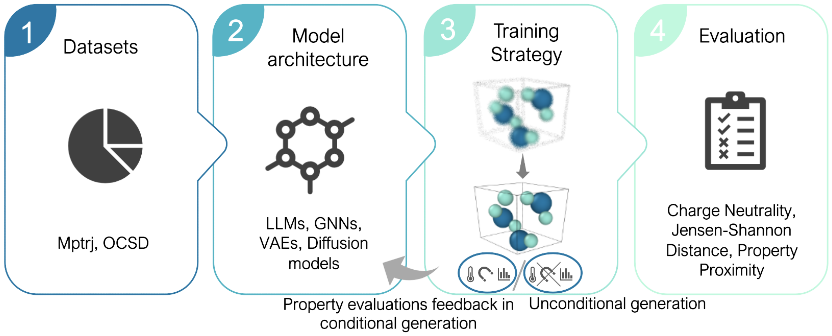

Figure 1: An overview of the generative AI paradigm for candidate structure generation and optimization that underpins much of the work reviewed herein.

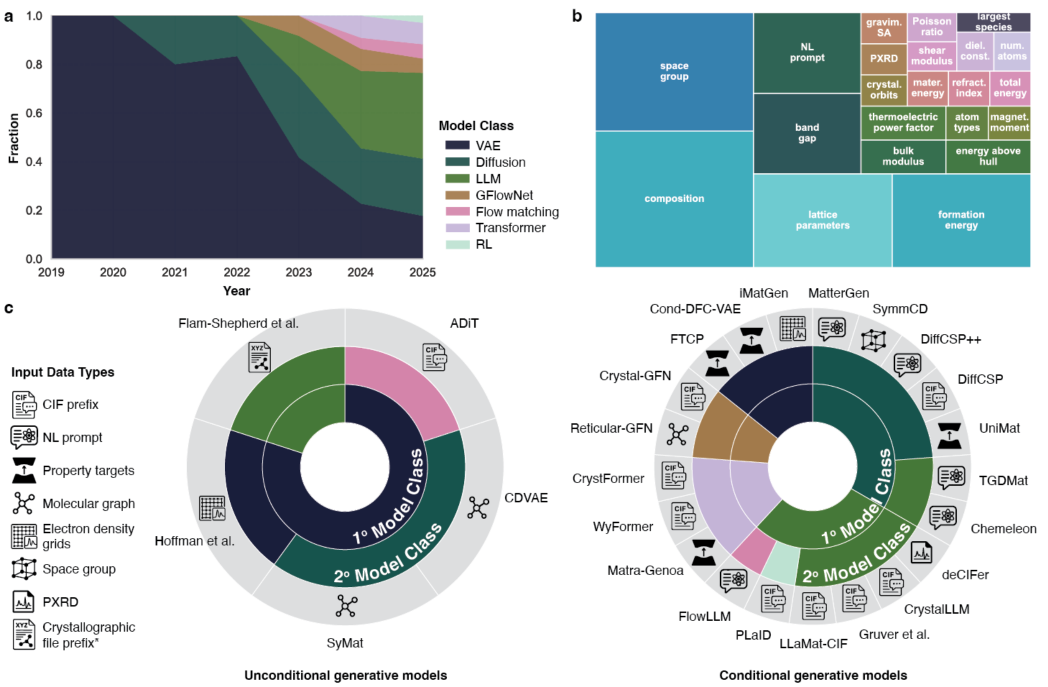

Figure 2: Overview of generative models for materials discovery discussed in this work. (a) Change over time of major model architectures discussed herein, showing early dominance of VAEs and the growth in prevalence of LLMs. (b) Treemap of target properties optimized across models; box size reflects the proportion of papers mentioning each property. Space group, composition, lattice parameters, and formation energy are the most common targets. (c) Pie charts illustrating the dominant model types used for unconditional (left) and conditional (right) materials generation, where the majority of conditional models can also do unconditional generation but not the other way around. The methods are clustered according to the primary (and, if applicable, secondary) model class. Colors match panel (a). Each model is annotated with its primary input data type; as the majority of current models return structures in CIF file format, this is not illustrated. *Abbreviations:* LLM = large language model; VAE = variational autoencoder; RL = reinforcement learning; NL prompt = natural language prompt; PXRD = powder X-ray diffraction. "CIF prefix" typically includes composition, space group, and lattice parameters; "Crystallographic file" refers to any file encoding structure data (e.g., XYZ, PDB, CIF).

A Desired Properties of a Crystal Generation Benchmark

Benchmarking plays a vital role in addressing this gap. Beyond enabling rigorous cross-model comparisons, it helps define what "good models" should look like in this rapidly evolving space. They offer reference points for assessing progress, provide structure for evaluating emerging methods, and help researchers, especially newcomers, understand how to design generative models with real-world impact.

Here, we list the desirable properties of the benchmark for crystal generation.

- **End-to-end automation with standardized evaluation.** For leaderboards and extensive evaluations across increasing new models, evaluations must run automatically across multiple datasets. The benchmark should provide automated structure validation, stability calculations using MLIPs, and property assessment without human intervention, enabling continuous maintenance of the leaderboard and seamless evaluation for users.
- **Expert validation of reference datasets and metrics.** Manual curation by crystallographers and materials scientists is essential to ensure the reference dataset (for instance, LeMat-Bulk, in this case) is free from duplicates, unstable structures, and annotation errors. Expert validation should also verify that evaluation metrics (fingerprinting, convex hull calculations) accurately capture physical and chemical plausibility.
- **Compatible with diverse model architectures.** The benchmark must accommodate different generative paradigms (VAEs, diffusion models, GFlowNets, LLMs, flow matching) and various crystal representations (CIF files, fractional coordinates, voxel grids, graph structures). The evaluation framework should accept any valid crystal structure format (or most of the widely used formats) as input.
- **Usable with black-box generative systems.** Many relevant systems are proprietary or use complex multi-stage pipelines. The benchmark should operate solely on generated crystal structures (the final CIF or structural files) without requiring access to model weights, latent representations, or intermediate outputs.
- **Probing capabilities beyond basic structure generation.** Real-world materials discovery requires more than generating valid crystals. The benchmark must evaluate conditional generation (property-targeted design), multi-objective optimization, synthesis constraints, and the ability to navigate complex structure-property relationships, not just unconditional sampling.
- **Cover diverse material systems and chemical spaces.** Materials science spans inorganics, organics, metals, semiconductors, and complex compounds across the periodic table. The benchmark should evaluate performance across different crystal systems, space groups, bonding types, and compositional complexity to assess true generalization capability.
- **Cover diverse materials design skills.** Holistic evaluation requires assessing multiple competencies: thermodynamic reasoning (stability prediction), chemical intuition (reasonable bonding), crystallographic knowledge (symmetry constraints), and inverse design capabilities (property-to-structure mapping).
- **Cover a range of generation difficulty levels.** To provide continuous improvement signals, the benchmark should span from simple binary compounds to complex multi-component systems, from high-symmetry to low-symmetry structures, and from well-studied to novel chemical spaces.
- **Impossible to completely solve with current models.** The benchmark should include challenging scenarios that push model limits: generating stable materials in unexplored chemical spaces, satisfying multiple competing constraints simultaneously, and discovering genuinely novel crystal structures that extend beyond training distributions.
- **Bridge computational prediction with experimental reality.** Unlike purely computational benchmarks, crystal generation must ultimately connect to synthesizable materials. The evaluation should incorporate synthesizability proxies, experimental validation pathways, and metrics that correlate with real-world materials discovery success.

B Evaluation metrics for materials generation

B.1 Unconditional Generation

Unconditional generation refers to the task of producing valid, stable crystal structures without targeting specific properties or constraints. The following metrics assess the fundamental quality of generated structures:

Fundamental Validity Metrics. These ensure the outputs are physically meaningful and chemically plausible. In different terms, they serve as a sanity check both for model development and inference time. Note that all metrics may not be relevant for every material system.

- **Charge Neutrality:** The total valence charge of all atoms must sum to zero:

  $$\sum_{i=1}^{N} q_i = 0 \tag{1}$$

  where q_i is the nominal oxidation state of atom i in the structure. For this to be calculated, the oxidation states of every atom in the structure must first be assigned. Here, we have developed a hierarchical structure for determining oxidation states and charge neutrality:

  1. If all atoms are metals, each atom is assigned a nominal oxidation state of zero and the structure is labeled as charge balanced. This metallic assignment is determined by using the Pymatgen “BVAnalyzer" class and the SMACT “metallicity-score" function with a cutoff of 0.7. If either of these are satisfied, the structure is labeled as a metal and therefore labeled charge balanced.
  2. If all atoms are not metals, the Pymatgen “get-oxi-state-decorated-structure" function Ong et al. [2013] is used to assign oxidation states and determine charge balance.
  3. However, the function used above can fail to find oxidation states for structures that are not well optimized. It is still necessary to determine whether these structures are charged balanced, particularly in the case of generative model benchmarks, when many structures may be too far from typical structures for the Pymatgen functions to analyze them. Here, we determine charge neutrality using a data driven approach from LeMatBulk Siron et al. [2025]. First, this workflow determines all the possible charge balanced compositions of oxidation states based on the observed oxidation states in LeMatBulk or the ICSD[Zagorac et al., 2019]. The most likely oxidation state assignments for this particular composition is determined by assigning each composition a score based on how probable that particular oxidation state configuration is, as determined by the distribution of oxidation states seen in LeMatBulk/ICSD. This score is determined by taking the geometric mean of all probabilities for each individual oxidation state. The geometric mean was chosen to normalize the score to the number of atoms in the structure, so scores between structures with two atoms and structures with ten atoms are comparable. Under the current implementation, if a composition exists from ICSD or LeMatBulk oxidation states which charge balances the structure, it passes the validity test. However, it is also possible to introduce a threshold here to filter out structures which require a combination of oxidation states which are highly uncommon, if a stricter validity metric is desired.
  4. If no charge balanced composition can be made using oxidation states from LeMatBulk/ICSD, the structure is passed to the next stage of the validity assessment. In this stage, all historically observed oxidation states for each element, including highly irregular ones, are used to see if the structure can be charge balanced. This list of oxidation states is drawn from Pymatgen’s “Element" class. Since this list includes some very irregular oxidation states (negative charges on metals, for example) an additional check is used to determine whether this charge balanced structure is viable. Since we don’t have access to probabilities for these very uncommon states, the oxidation state of each element in the system is correlated to the Pauli electronegativity of that element. A negative correlation is desired, as this would show increasing electronegativity leads to a decreasing oxidation state. Any structures which show a positive correlation are labeled invalid, as this would reflect a very chemically implausible structure overall.

- **Minimum Interatomic Distance:** All interatomic distances d_{ij} must exceed a cutoff value $d_{\min}$ to prevent atomic overlap. We suggest adopting 0.5 Å.

$$d_{ij} > d_{\min} \quad \forall\, i \neq j \tag{2}$$

Mass density and atomic number density : are within reasonable ranges. Mass density is given by $\rho = \frac{M_{\text{total}}}{V_{\text{cell}}}$, in (g/cm^3). The latter is expressed in atoms/Å^3. We take upper bounds of 25 g/cm^3 and 0.5 atoms/Å^3, respectively.

Valid crystallographic representation : a good proxy is to determine whether a structure is CIF-readable using *pymatgen*.

Lattice Parameters : are within reasonable ranges. We take upper bounds of 100 Åfor a,b,c and 180 degrees for α, β, γ respectively, and lower bounds of 1 Åand zero degrees for a,b,c and α, β, γ, respectively.

Stability metrics. These assess the thermodynamic and energetic properties of generated structures:

- **Formation Energy (E_f):**

$$E_f = E_{\text{tot}}(\text{compound}) - \sum_i n_i \mu_i \tag{3}$$

where E_{tot} is the total energy of the crystal, n_i is the number of atoms of element i, and μ_i is the chemical potential of the pure element. The result is normalized per atom: $E_f^{\text{per atom}} = \frac{E_f}{\sum_i n_i}$. We want it to be as small (and negative) as possible.

The chemical potentials μ_i are derived from the LeMaterial-Bulk dataset by taking the minimum energy among all single-element structures for each element: $\mu_i = \min_{k \in S_i} \left(E_{\text{norm}}^{(k)} \right)$ where S_i is the set of all single-element structures containing element i.

Multi-MLIP Ensemble Implementation: The formation energy metric supports ensemble statistics across multiple MLIPs (ORB, MACE, UMA). For each structure, ensemble statistics are computed as:

$$\langle E_f \rangle = \frac{1}{N_{\text{MLIP}}} \sum_{k=1}^{N_{\text{MLIP}}} E_f^{(k)} \tag{4}$$

$$\sigma_{E_f} = \sqrt{\frac{1}{N_{\text{MLIP}} - 1} \sum_{k=1}^{N_{\text{MLIP}}} \left(E_f^{(k)} - \langle E_f \rangle \right)^2} \tag{5}$$

where $E_f^{(k)}$ is the formation energy predicted by the k-th MLIP. The implementation extracts pre-computed ensemble statistics from structure properties (`formation_energy_mean`, `formation_energy_std`) or calculates them from individual MLIP results (`formation_energy_orb`, `formation_energy_mace`, etc.). A minimum of 2 MLIPs is required for ensemble statistics.

- **Energy Above Convex Hull (E_{hull}):**

$$E_{\text{hull}} = E_{\text{tot}} - E_{\text{hull}}^{\min} \tag{6}$$

Structures with $E_{\text{hull}} \leq 0$ are considered stable, while values below approximately 0.1 eV/atom are often deemed metastable. We take LeMat-Bulk [Siron et al., 2024] as reference point for calculating the convex hull.

The convex hull is constructed by filtering the LeMat-Bulk dataset to include only compounds containing elements present in the target composition, creating `PDEntry` objects, and using Pymatgen's `PhaseDiagram.get_decomp_and_e_above_hull()` method. The implementation handles charged species by extracting neutral elements before phase diagram construction. Multi-MLIP ensemble statistics follow the same formulation as formation energy: $\langle E_{\text{hull}} \rangle = \frac{1}{N_{\text{MLIP}}} \sum_{k=1}^{N_{\text{MLIP}}} E_{\text{hull}}^{(k)}$ with corresponding standard deviation calculations.

- **Relaxation Stability:** Use an ensemble of Machine Learning Interatomic Potentials to relax the generated structures (each one is done independently). Then, compute the Root Mean Square Deviation (RMSD) between pre- and post-relaxation atomic positions:

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{r}_i^{\text{init}} - \mathbf{r}_i^{\text{relax}}\|^2} \tag{7}$$

  Low RMSD indicates minimal distortion and structural robustness under optimization.

  The implementation calculates individual RMSD values for each MLIP relaxation, then computes ensemble statistics: $\langle\text{RMSD}\rangle = \frac{1}{N_{\text{MLIP}}}\sum_{k=1}^{N_{\text{MLIP}}}\text{RMSD}^{(k)}$ where $\text{RMSD}^{(k)}$ is the relaxation RMSD from the k-th MLIP. The metric extracts pre-computed values from structure properties (`relaxation_rmsd_mean`, `relaxation_rmsd_std`) or calculates ensemble statistics from individual MLIP results (`relaxation_rmsd_orb`, `relaxation_rmsd_mace`, etc.). Lower values indicate better structural stability under relaxation.

Novelty, Uniqueness, and Diversity Metrics. These evaluate how effectively a model explores the chemical space:

- **Novelty ($\mathcal{N}$):** Evaluates the fraction of generated structures that are not present in a reference dataset of known materials. The novelty score is defined as:

$$\mathcal{N} = \frac{|\{x \in G \mid x \notin T\}|}{|G|} \tag{8}$$

  where G is the set of generated structures and T is the reference dataset (LeMat-Bulk).

  The implementation supports two comparison methods: **BAWL fingerprinting** using crystallographic hash strings with Weisfeiler-Lehman graph kernels, and **structure matching** using Pymatgen's symmetry-aware structural comparison algorithms. For BAWL, novelty is determined by checking if the generated structure's fingerprint exists in the pre-computed reference fingerprint set. For structure matching, each generated structure is compared against reference structures with overlapping elemental compositions using space group analysis and atomic position matching with configurable tolerances. In our paper, we report results using the structure matcher approach for more robust structural comparison against the LeMat-Bulk reference dataset.

- **Uniqueness ($\mathcal{U}$):** Measures the fraction of unique structures within the generated set to assess internal diversity. The uniqueness score is defined as:

$$\mathcal{U} = \frac{|\text{unique}(G)|}{|G|} \tag{9}$$

  where $\text{unique}(G)$ returns the set of unique structures based on their fingerprints.

  The metric is implemented as a structure-level continuous scoring system rather than binary classification. For BAWL fingerprinting, individual uniqueness scores are assigned as $u_i = 1/c_i$, where c_i is the count of structures sharing the same fingerprint within the generated set. This assigns a score of 1.0 to truly unique structures while proportionally penalizing duplicated structures. For structure matching, the implementation uses pairwise comparison with an ordered approach: structure i is considered unique if it is not equivalent to any structure j where $j < i$, ensuring deterministic selection of the first occurrence as the unique representative. The overall uniqueness metric is computed as $\mathcal{U} = \frac{1}{|G|}\sum_{i=1}^{|G|} u_i$. Both BAWL fingerprinting and structure matching methods are supported, with structure matching used for paper results.

- **S.U.N. and M.S.U.N. Rates:** Proportion of generated structures that are simultaneously Stable (or Metastable), Unique, and Novel:

$$\text{S.U.N. Rate} = \frac{|\{x \in G \mid E_{\text{hull}}(x) \leq 0,\ x \notin T,\ x \text{ is unique}\}|}{|G|} \tag{10}$$

$$\text{M.S.U.N. Rate} = \frac{|\{x \in G \mid 0 < E_{\text{hull}}(x) \leq \tau,\ x \notin T,\ x \text{ is unique}\}|}{|G|} \tag{11}$$

where τ is a metastability threshold (commonly 0.08-0.1 eV/atom, though this varies across studies [Miller et al., 2024, Gruver et al., 2024, Zeni et al., 2025]).

The implementation follows a hierarchical computation order: **Stability** $\rightarrow$ **Uniqueness** $\rightarrow$ **Novelty**. First, structures are classified as stable ($E_{\text{hull}} \leq 0$) or metastable ($0 < E_{\text{hull}} \leq \tau$) using energy above hull values computed by the Multi-MLIP stability preprocessor. Then, uniqueness is evaluated within each stability class separately using the chosen comparison method. Finally, novelty is assessed for unique structures from each stability class. This hierarchical approach provides detailed metrics at each evaluation stage: stability counts, unique-within-stable/metastable counts, and final SUN/MSUN counts. The Multi-MLIP preprocessor assigns ensemble stability properties (e.g., `e_above_hull_mean`) to structure objects, enabling robust stability classification across multiple MLIPs (ORB, MACE, UMA). We set τ to 0.1 eV/atom for assembling results.

- **Diversity:** plot the Distribution analysis of space groups, elemental compositions, and lattice parameters in comparison to reference datasets. But also:
  - Composition, Space Group, Lattice and Atomic Site Entropy: Suppose you generated N structures, and you count the frequency f_i of each element i (e.g., O, Fe, Zn…) across all structures. Normalize to get a probability distribution: $p_i = \frac{f_i}{\sum_j f_j}$. Then compute Shannon entropy: $H = -\sum_i p_i \log p_i$ and the Vendi Score [Friedman and Dieng, 2022], which is the exponential of the Shannon Entropy. The above example is for composition entropy, but this methodology is also applied to the other criteria listed above in our diversity benchmark.

Distribution-Level Metrics. When trying to measure how well the distribution of generated structures matches the real material distribution, we can use:

- **Jensen-Shannon Distance** [Fuglede and Topsoe, 2004]:

$$\text{JSD}(P, Q) = \sqrt{\frac{1}{2} D_{KL}(P||M) + \frac{1}{2} D_{KL}(Q||M)} \tag{12}$$

  where P and Q are distributions of generated and real samples, M is the average of the two distributions ($\frac{1}{2}(P+Q)$), and D_{KL} is the Kullback Leibler divergence.

- **Maximum Mean Discrepancy (MMD)** [Tolstikhin et al., 2016]:

$$\text{MMD}^2(P, Q) = \mathbb{E}_{x,x'}[k(x, x')] + \mathbb{E}_{y,y'}[k(y, y')] - 2\mathbb{E}_{x,y}[k(x, y)] \tag{13}$$

  where P and Q are distributions of generated and real samples, and k is a kernel function.

- **Fréchet Distance Metrics** [Heusel et al., 2017, Preuer et al., 2018]: Adaptations like Fréchet ChemNet Distance (FCD) compare the distributions of generated and reference structures:

$$\text{FD}(G, T) = \|\mu_G - \mu_T\|^2 + \text{Tr}\left(\Sigma_G + \Sigma_T - 2(\Sigma_G \Sigma_T)^{1/2}\right) \tag{14}$$

  where μ and Σ represent the mean and covariance of embeddings.

Model Efficiency This measures how effectively a model learns from limited training data [Gao et al., 2022]:

- **Generic metrics**: training dataset size, number of model parameters, number of epochs required for training, training time and associated computational infrastructure, inference time on 10k structures.
- **Learning Curve Analysis:** Performance (e.g., S.U.N. rate, property prediction accuracy) as a function of the number of expensive function evaluations (e.g., DFT calculations) required for training, i.e., the number of labeled data points.

Herfindahl-Hirschman Index (HHI) Metrics. The Herfindahl-Hirschman Index quantifies supply risk concentration for materials by measuring the concentration of element production sources and reserves. For a given crystal structure with composition, we compute:

- **Compound HHI Value**: For a compound with chemical formula represented by composition C:

$$\text{HHI}_{\text{compound}} = \sum_i x_i \cdot \text{HHI}_i \tag{15}$$

  where x_i is the fractional composition of element i in the compound, and HHI_i is the element-specific HHI value.
- **Production HHI**: Measures supply risk based on concentration of element production sources (market concentration):

$$\text{HHI}_{\text{production}} = \sum_j s_j^2 \times 10000 \tag{16}$$

  where s_j is the market share of producer j for a given element.
- **Reserve HHI**: Measures long-term supply risk based on concentration of element reserves (geographic distribution):

$$\text{HHI}_{\text{reserve}} = \sum_k r_k^2 \times 10000 \tag{17}$$

  where r_k is the fraction of global reserves held by country/region k.
- **Scaling Convention**: HHI values are typically scaled from the classical range [0, 10000] to a convenience range [0, 10]:

$$\text{HHI}_{\text{scaled}} = \frac{\text{HHI}_{\text{classical}}}{1000} \tag{18}$$

- **Combined HHI Score**: The final benchmark score combines both production and reserve metrics using weighted averaging:

$$\text{HHI}_{\text{combined}} = w_{\text{prod}} \cdot \text{HHI}_{\text{production}} + w_{\text{res}} \cdot \text{HHI}_{\text{reserve}} \tag{19}$$

  where $w_{\text{prod}} = 0.25$ and $w_{\text{res}} = 0.75$ by default, prioritizing long-term supply security over short-term market dynamics.
- **Missing Element Handling**: Elements not found in the HHI lookup tables are assigned the maximum risk value (10000 unscaled / 10 scaled) to represent maximum supply uncertainty for rare or untracked elements.
- **Risk Categories**: For the scaled [0, 10] range:

$$\text{Low Risk} : \text{HHI}_{\text{scaled}} \leq 2.0 \tag{20}$$

$$\text{Moderate Risk} : 2.0 < \text{HHI}_{\text{scaled}} \leq 5.0 \tag{21}$$

$$\text{High Risk} : \text{HHI}_{\text{scaled}} > 5.0 \tag{22}$$

B.2 Conditional Generation

Conditional generation involves producing crystal structures that satisfy specific constraints or exhibit targeted properties. Evaluating such models requires metrics that assess both adherence to conditions and overall structural quality.

Property Targeting Metrics. These measure how well generated structures match specified target properties:

- **Top-k** values: compute the mean and standard of top-k property values, for $k = 1, 10, 100$, that maximize or minimize an objective for generated material structures.
- **Property Proximity:** The deviation between the target property value p_{target} and the achieved value $p_{\text{generated}}$:

$$\text{Error}(p) = |p_{\text{generated}} - p_{\text{target}}| \tag{23}$$

- **Success Rate:** Fraction of generated structures whose properties fall within an acceptable range around the target:

$$\text{Success Rate} = \frac{|\{x \in G \mid |p(x) - p_{\text{target}}| \leq \delta\}|}{|G|} \tag{24}$$

  where δ is the tolerance threshold.
- **Conditional S.U.N. Rate:** Proportion of stable, unique, and novel structures that also meet the conditional property constraints. Additionally, we calculate the V.S.U.N. rate, which also includes whether the structures pass our validity benchmarks.

Constraint Adherence Metrics. These evaluate how well generated structures conform to specified structural constraints:

- **Space Group Fidelity:** For symmetry-conditioned generation, the proportion of structures that correctly exhibit the specified space group as defined by Pymatgen's SpacegroupAnalyzer.
- **Composition Fidelity:** For composition-conditioned generation, the accuracy of incorporating specified elements in the correct stoichiometries.
- **Wyckoff Position Accuracy:** For models conditioning on crystallographic sites, the correctness of atom placement according to specified Wyckoff positions [Kazeev et al., 2025].

Multi-Objective Optimization Metrics. These assess models tasked with optimizing multiple properties simultaneously:

- **Pareto Optimality:** Analysis of the non-dominated solutions in the multi-dimensional property space.
- **Hypervolume Indicator:** The volume of the dominated portion of the objective space, relative to a reference point.
- **MOQD Score:** Quality-diversity metric that rewards finding diverse sets of high-performing solutions across different feature dimensions [Janmohamed et al., 2024].

B.3 Going further

While our benchmark focuses on core objectives such as Conditional S.U.N, diversity, validity, we recognize the importance of additional evaluation axes that capture real-world utility. Metrics assessing *out-of-distribution generalization*—including extrapolation to unseen chemistries, scalability to larger systems, and rediscovery of held-out targets—are critical for assessing the robustness and true generative capabilities of models. Similarly, *synthesizability assessment* metrics such as synthetic accessibility scores, retrosynthetic success rates, or proximity to known materials offer insight into the practical feasibility of generated candidates. These aspects, though not included in this release, represent essential directions for future benchmarking and method development.

Standardizing Convex Hull Computation and Stability To make stability a trustworthy benchmark for generative crystal design, E_{hull} must be built with fully disclosed and identical DFT settings. Because E_{hull} measures the distance of a structure's formation energy from the multiphase convex hull, its value changes with every additional phase; therefore, authors should always disclose the full DFT workflow (functional, U values, k-mesh, energy corrections) and the total number of DFT-relaxed formation energies that define the hull. Values derived from spaces with fewer than two competing phases should be flagged as unreliable. Machine-learning interatomic potentials are convenient for screening [Duval et al., 2023b, 2024] but systematically under-estimate E_{hull} [Nong et al., 2025], so MLIP-based hulls must be recalibrated with consistent first-principles data before being used for benchmarking. Additionally, E_{hull} reflects thermodynamic stability only at 0K and 0atm, so kinetic stability must be verified separately–for example, by ensuring that phonon spectra contain no imaginary modes. Finally, the common "$\leq$ 0 meV" criterion should be applied cautiously: numerous compounds synthesized in the laboratory sit 50–150 meV per atom above the 0K hull, highlighting the need to augment databases with additional, consistently computed DFT polymorphs to improve phase-diagram fidelity and to contextualise what constitutes a realistically synthesizable region.

Out-of-Distribution Generalization These metrics specifically target the model's ability to generate valid structures in previously unexplored regions:

- **Extrapolation Success:** Performance on generating structures with elements, stoichiometries, or structure types not seen during training.
- **Size Generalization:** Ability to generate larger or more complex structures than those in the training set.
- **Rediscovery Rate:** Ability to generate known high-performance materials that were explicitly excluded from training, demonstrating the model's capacity to learn fundamental design principles rather than merely memorizing training examples.

Synthesizability Assessment These metrics evaluate the practical realizability of generated structures:

- **Synthetic Accessibility Score:** Heuristic metrics adapted from drug discovery, such as SAscore [Seo et al., 2024], that estimate synthetic feasibility based on structural complexity or similarity to known materials.
- **Retrosynthesis Success Rate:** The proportion of generated structures for which computational retrosynthesis tools like AiZynthFinder [Guo and Schwaller, 2025] or ASKCOS [Gao et al., 2024] can identify plausible synthetic pathways.
- **Proximity to Synthesized Materials:** Distance in feature space or embedding space to the nearest experimentally synthesized structure.

C Benchmark workflow and results

The benchmark evaluation follows a structured two-phase workflow designed to ensure computational efficiency and meaningful comparison by operating only on structurally valid materials. The workflow enforces a mandatory validity filtering step followed by selective preprocessing and evaluation phases.

C.1 Phase 1: Mandatory Validity Assessment and Filtering

Input Processing: LeMat-GenBench accepts input structures from multiple sources: (1) individual CIF file paths in text format, (2) directories containing CIF files processed recursively, or (3) CSV files containing structures in various formats (JSON dictionaries, CIF strings, or pymatgen Structure objects).

Validity Benchmark Execution: All input structures are subjected to the standardized validity criteria described in Section 3.1 (cf. Validity). The `ValidityBenchmark` applies these checks uniformly and reports aggregate validity rates, failure mode distributions, and structural property statistics.

Validity Preprocessing: In parallel, the `ValidityPreprocessor` attaches validity metadata to each structure, assigns unique identifiers, and generates detailed validation reports to ensure traceability between submitted inputs and benchmark results.

Critical Filtering Step: Only structures passing all validity checks are retained for downstream benchmarks. This step reduces computational overhead for expensive operations (e.g., MLIP calculations) and ensures that evaluation metrics reflect realistic material properties rather than artifacts of invalid structures. Filtering outcomes are comprehensively logged for transparency.

C.2 Phase 2: Selective Preprocessing and Benchmark Evaluation

Preprocessor Configuration: Based on the selected benchmark families, the system automatically determines required preprocessing steps. The configuration logic maps benchmark requirements to preprocessors: fingerprint-based benchmarks (novelty, uniqueness, SUN) require `FingerprintPreprocessor` for BAWL/short-BAWL methods, distribution-based benchmarks require `DistributionPreprocessor`, and stability assessments require `MultiMLIPStabilityPreprocessor`. All preprocessors attach their computed outputs as attributes within the `properties` dictionary of each pymatgen `Structure` object, enabling seamless data flow between preprocessing and benchmark evaluation phases while maintaining full traceability of computed features.

Fingerprint Preprocessing: When fingerprint-based evaluation is required, the `FingerprintPreprocessor` computes structural fingerprints using the specified method (BAWL, short-BAWL [Siron et al., 2025], or PDD [Widdowson and Kurlin, 2021]). This preprocessor is bypassed entirely when structure-matcher is selected as the fingerprinting method, since structure-matcher performs direct pairwise structural comparison using pymatgen's `StructureMatcher` algorithm rather than pre-computed fingerprints. The structure-matcher approach uses configurable tolerance thresholds (default: 0.1) to determine structural equivalence through lattice parameter matching, atomic position comparison, and symmetry analysis, providing more rigorous but computationally expensive structural comparison than hash-based fingerprinting methods.

Distribution Preprocessing: For benchmarks requiring compositional or structural distribution analysis, the `DistributionPreprocessor` computes statistical descriptors needed for Maximum Mean Discrepancy (MMD) and Jensen-Shannon divergence calculations. This preprocessor extracts compositional features, structural parameters, and other distributional characteristics required for comparing generated structures against reference databases.

Multi-MLIP Preprocessing: The `MultiMLIPStabilityPreprocessor` performs the most computationally intensive preprocessing, utilizing multiple machine learning interatomic potentials (MLIPs) including ORB v3[Rhodes et al., 2025], MACE-MP[Batatia et al., 2023], and UMA[Wood et al., 2025]. This preprocessor performs: (1) structure relaxation using configurable force convergence criteria (default: 0.02 eV/Å), (2) formation energy calculations against reference states, (3) energy above hull computations using convex hull analysis, and (4) MLIP embedding extraction for Fréchet distance calculations.

Benchmark Execution: Following preprocessing, the system executes selected benchmarks on the processed valid structures. Each benchmark operates independently with dedicated memory management and error handling. The execution order is optimized to minimize memory conflicts, with computationally expensive benchmarks (multi-MLIP stability) scheduled with aggressive memory cleanup between operations. The benchmark system generates comprehensive JSON output containing: (1) run metadata including structure counts, benchmark configurations, and execution timestamps, (2) validity filtering metadata tracking the transition from input structures to valid structures, (3) detailed results for each benchmark family with appropriate statistical summaries, and (4) pre-processor results and intermediate data for reproducibility and debugging. Further information on metrics and their implementation is available in Appendix B.

Table 3: Core generation metrics: validity, uniqueness, novelty, energy, stability, and metastability. Arrows indicate optimization direction. Bold indicates best performance, underlined indicates second-best. Models below the separator (MatterGen, PLaID++, WyFormer, and WyFormer-DFT) incorporate structural relaxation in their generation or evaluation pipelines.

Model	**#**	**Validity**				**Unique ↑**	**Novel ↑**	**Energy-based ↓**			**Stability ↑**			**Metastability ↑**		
	Structures	**Valid ↑**	**CN ↑**	**MinDist ↑**	**PhysPlau ↑**			**FormE (Std)**	**E_{hull} (Std)**	**RMSD (Std)**	**Stable**	**U-Stable**	**SUN**	**Metastable**	**U-Meta**	**MSUN**
ADiT	1000	905	$\underline{989}$	914	**1000**	897	277	-0.663 ± 0.910	0.412 ± 0.544	0.393 ± 0.408	$\underline{25}$	24	2	196	195	5
Crystalformer	1000	699	963	729	$\underline{999}$	694	318	-0.169 ± 1.482	0.745 ± 1.350	0.660 ± 1.066	16	16	6	191	188	14
DiffCSP	1000	961	964	997	**1000**	$\underline{958}$	$\underline{666}$	$\underline{-0.667 \pm 0.828}$	0.317 ± 0.294	0.596 ± 0.642	19	19	**12**	188	187	32
DiffCSP++	1000	957	957	**1000**	**1000**	956	652	-0.535 ± 0.862	0.472 ± 0.458	0.732 ± 0.829	23	23	$\underline{11}$	135	134	22
LLaMat2	1000	860	974	885	997	848	331	-0.471 ± 1.056	0.472 ± 0.790	0.525 ± 0.663	22	22	6	$\underline{218}$	$\underline{211}$	16
SymmCD	1000	723	961	750	**1000**	722	476	-0.017 ± 1.184	0.922 ± 1.048	0.893 ± 1.065	13	13	5	117	116	11
MatterGen	1000	$\underline{962}$	963	$\underline{999}$	**1000**	**961**	**700**	$\mathbf{-0.715 \pm 0.790}$	$\underline{0.236 \pm 0.242}$	$\underline{0.372 \pm 0.424}$	$\underline{25}$	$\underline{25}$	$\underline{11}$	204	204	**66**
PLaID++	1000	**985**	**991**	993	$\underline{999}$	873	241	-0.495 ± 0.465	$\mathbf{0.130 \pm 0.225}$	$\mathbf{0.119 \pm 0.242}$	**29**	**27**	3	**519**	**431**	$\underline{51}$
WyFormer	1000	938	950	987	**1000**	938	650	-0.474 ± 0.928	0.523 ± 0.495	0.750 ± 0.790	18	18	7	91	91	7
WyFormer-DFT	1000	951	952	**1000**	$\underline{999}$	946	662	-0.646 ± 0.882	0.349 ± 0.363	0.405 ± 0.611	17	17	10	178	174	32

Table 4: Training datasets and data sources used for the reported generative crystal structure models

Model	Training Dataset	Source of Submitted Structures
ADiT	MP-20	Authors of [Joshi et al., 2025]
Crystalformer	MP-20	Figshare of [Kazeev et al., 2025][3]
DiffCSP	MP-20	Figshare of [Kazeev et al., 2025][3]
DiffCSP++	MP-20	Figshare of [Kazeev et al., 2025][3]
LLaMat2	MP-20	Authors of [Mishra et al., 2024]
MatterGen	MP-20	Figshare of [Kazeev et al., 2025][3]
PLaID++	MP-20	Authors of [Xu et al., 2025a]
SymmCD	MP-20	Figshare of [Kazeev et al., 2025][3]
WyFormer-DiffCSP++	MP-20	Authors of [Kazeev et al., 2025]
WyFormer-DiffCSP++-DFT	MP-20	Authors of [Kazeev et al., 2025]

---

[3] `https://figshare.com/articles/dataset/Generated_crystals_for_WyFormer_DiffCSP_DiffCSP_WyCryst_SymmCD_CrystalFormer_MiAD/29145101`

D Environmental and Sustainability Considerations

The application of generative models to materials discovery presents significant opportunities for advancing environmental sustainability goals. As global challenges related to climate change, resource depletion, and environmental degradation intensify, the need for novel materials with reduced environmental footprints becomes increasingly urgent. Generative approaches can accelerate the discovery of sustainable alternatives by explicitly incorporating environmental criteria into the design process.

One promising direction involves the targeted generation of materials with reduced reliance on critical or environmentally problematic elements. By conditioning generative models on compositional constraints that exclude toxic, rare, or environmentally harmful elements, researchers can guide exploration toward more sustainable regions of chemical space. Similarly, models can be trained to prioritize earth-abundant elements and avoid those associated with problematic extraction practices or geopolitical supply risks.

Energy-related applications represent another frontier where generative models could significantly impact sustainability outcomes. The discovery of more efficient catalysts for renewable energy production, improved battery materials for energy storage, and novel photovoltaic materials could accelerate the transition away from fossil fuels. By specifically targeting properties relevant to these applications, generative models can focus computational and experimental resources on high-impact sustainability domains.

Life-cycle considerations present a more complex but equally important target for integration with generative approaches. Ideally, materials should be designed not only for performance but also for recyclability, biodegradability, or other end-of-life scenarios that minimize environmental impact. Incorporating such considerations into generative frameworks remains challenging due to the complex, multi-faceted nature of life-cycle assessment, but represents a crucial direction for future research.

The computational efficiency of generative processes themselves also warrants consideration from a sustainability perspective. As models grow in complexity and scale, their energy consumption and carbon footprint increase accordingly. Developing more efficient architectures, training procedures, and sampling approaches could reduce the environmental impact of the discovery process itself, aligning computational means with environmental ends. This consideration becomes particularly important as generative approaches scale to industrial applications and high-throughput discovery platforms.

The ultimate success of generative approaches in advancing sustainability will depend not only on technical capabilities but also on intentional alignment with environmental objectives. By explicitly incorporating sustainability metrics into reward functions, objective functions, and evaluation criteria, the materials community can ensure that generative models contribute to addressing environmental challenges rather than merely accelerating traditional discovery paradigms without regard for sustainability implications.