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.

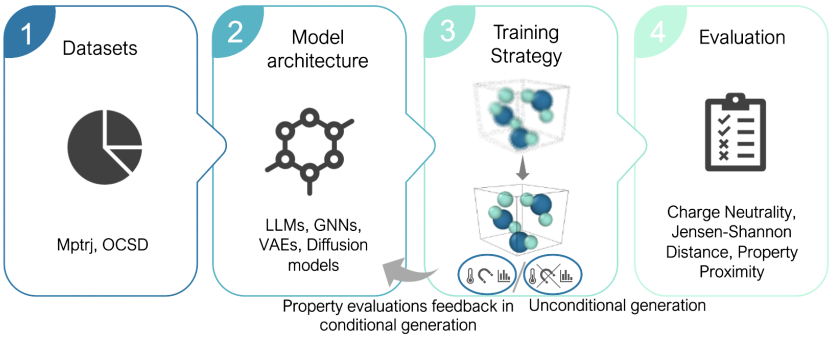

Figure 1: An overview of the generative AI paradigm for candidate structure generation and optimization that underpins much of the work reviewed herein.

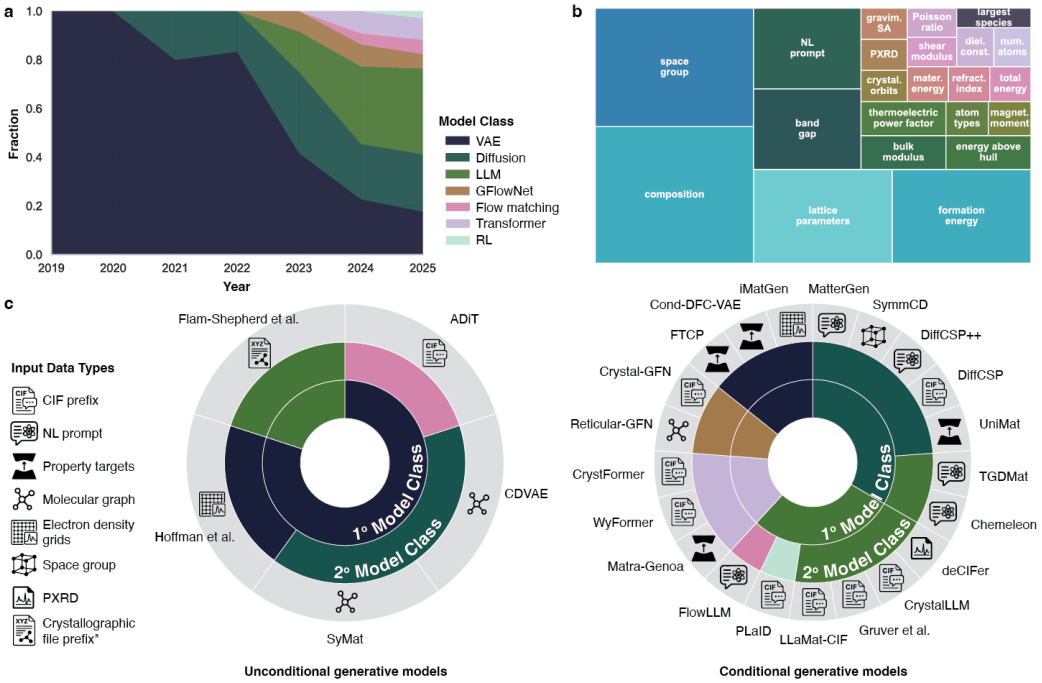

Figure 2: Overview of generative models for materials discovery discussed in this work. (a) Change over time of major model architectures discussed herein, showing early dominance of VAEs and the growth in prevalence of LLMs. (b) Treemap of target properties optimized across models; box size reflects 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.
  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. If no charge balanced composition can be made using these oxidation states, the structure is labeled invalid. The most likely oxidation state assignments for this particular composition, each composition is assigned a score based on how probable that particular oxidation state configuration is, as determined by the distribution of oxidation states seen in LeMatBulk. This score is determined by multiplying all of the probabilities for each individual oxidation state together and multiplying by the number of elements for a normalization. If the probability is greater than 0.001, the structure passes the validity test. Otherwise, to be charge balanced it requires a combination of oxidation states which are extremely rare, and therefore, is not valid.

- **Minimum Interatomic Distance:** All interatomic distances d_{ij} must exceed a cutoff value $d_{\min}$ to prevent atomic overlap. We suggest adopting 0.7 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/cm³ 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 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 "≤ 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) preprocessor results and intermediate data for reproducibility and debugging. Further information on metrics and their implementation is available in Appendix B.

Table 3: Model Evaluation Metrics

Model	#	Validity				Unique ↑	Novel ↑	Energy-based			Stability			Metastability			Distribution			Diversity				HHI	
	Structures	Valid ↑	CN ↑	MinDist ↑	PhysPlau ↑			FormE (Std) ↓	E_{hull} (Std) ↓	RMSD (Std) ↓	Stable ↑	U-Stable ↑	SUN ↑	Metastable ↑	U-Meta ↑	MSUN ↑	JS ↓	MMD ↓	FID ↓	ElemDiv ↑	SGDiv ↑	SizeDiv ↑	SiteDiv ↑	Prod ↓	Res ↓
ADiT [Joshi et al., 2025]	1000	812	882	914	1000	806	252	-2.288 ± 3.807	2.111 ± 4.418	0.389 ± 0.393	19	18	2	108	107	5	0.522	0.003	1.848	0.703	0.022	0.270	14.221	3.428	2.661
Crystalformer [Kazeev et al., 2024]	1000	577	687	642	796	572	247	-1.722 ± 9.741	2.728 ± 5.962	0.587 ± 0.858	13	13	4	106	104	5	0.273	0.003	2.489	0.695	0.313	0.322	17.385	3.830	2.785
DiffCSP [Jiao et al., 2024]	1000	732	733	823	825	729	475	-2.353 ± 3.730	1.706 ± 4.224	0.519 ± 0.622	17	17	**11**	109	108	18	0.464	0.007	1.796	0.695	0.104	0.279	14.277	3.420	2.628
DiffCSP++ [Kazeev et al., 2025]	1000	748	748	858	858	747	462	-4.308 ± 7.771	2.701 ± 5.585	0.661 ± 0.776	20	20	10	87	86	15	0.243	0.005	2.387	0.686	0.391	0.307	20.007	3.535	2.692
LLaMat2 [Mishra et al., 2024]	1000	779	873	885	997	769	286	-1.120 ± 4.707	2.572 ± 5.673	0.487 ± 0.617	21	21	6	125	122	11	0.329	0.003	1.431	0.703	0.187	0.269	9.153	3.994	2.988
MatterGen [Zeni et al., 2025]	1000	739	740	829	830	738	499	-2.218 ± 2.806	1.731 ± 4.184	0.334 ± 0.399	19	19	10	136	136	**42**	0.439	0.006	1.798	0.644	0.126	0.276	12.109	3.525	2.650
PLaID++ [Lee et al., 2025]	1000	**960**	965	993	999	**848**	228	-2.325 ± 2.994	3.452 ± 6.161	0.114 ± 0.240	**25**	**24**	7	218	183	26	0.446	0.035	3.008	0.652	0.204	0.238	5.948	5.246	3.394
SymmCD [Kazeev et al., 2025]	1000	561	737	642	861	560	343	-1.161 ± 8.279	2.816 ± 5.505	0.763 ± 0.965	9	9	3	64	64	3	0.236	0.006	1.879	0.703	0.378	0.320	18.088	3.549	2.692
WyFormer-DiffCSP++ [Kazeev et al., 2025]	1000	798	810	987	1000	798	530	-3.565 ± 8.384	2.048 ± 5.338	0.722 ± 0.794	16	16	6	70	70	6	0.238	0.008	1.436	0.695	0.370	0.309	21.638	3.601	2.701
WyFormer-DFT [Kazeev et al., 2025]	1000	839	839	1000	999	834	**569**	-4.749 ± 7.717	2.141 ± 5.664	0.380 ± 0.609	15	15	9	128	124	25	0.271	0.011	2.129	0.712	0.387	0.302	21.900	3.495	2.666

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., 2025]
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]

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

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

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.