LeMat-GenBench: A Unified Evaluation Framework for Crystal Generative Models

Siddharth Betala [1] **Samuel P. Gleason** [1] **Ali Ramlaoui** [1] **Andy Xu** [2] **Georgia Channing** [3] **Daniel Levy** [4] **Clementine Fourrier** [3] **Nikita Kazeev** [5] **Chaitanya K. Joshi** [6] **Sekou-Oumar Kaba** [4] **Felix Therrien** [4] **Alex Hernandez-Garcia** [4] **Rocio Mercado** [7] **N. M. Anoop Krishnan** [8] **Alexandre Duval** [1]

[1]*Entalpic* [2]*Harvey Mudd College* [3]*Hugging Face* [4]*Mila* [5]*National University of Singapore* [6]*University of Cambridge* [7]*Chalmers University of Technology,* [8]*Indian Institute of Technology Delhi*. Correspondence to: Siddharth Betala siddharth.betala@entalpic.ai, Alexandre Duval alexandre.duval@entalpic.ai.

1. Introduction

The discovery of inorganic crystalline materials underpins advances across energy storage, catalysis, electronics, and structural materials, making it central to both fundamental science and technological innovation [1, 2]. Historically, such discovery has relied on an Edisonian loop of expert intuition and experimental validation [3], occasionally accelerated by first-principles simulations such as density functional theory (DFT) [4, 5]. While DFT provides reliable insights into structural stability and properties, its computational cost and reliance on fully specified atomic configurations limit its applicability for large-scale exploration of chemical space.

Advances in ML [6, 7], particularly geometric graph neural networks [8] and generative models [9, 10, 11, 12, 13, 14], have enabled the direct generation of candidate crystal structures. However, despite rapid progress, the lack of standardized and reproducible evaluation protocols has made it difficult to assess whether new models meaningfully advance discovery or merely exploit evaluation artifacts.

Existing studies differ substantially in how they define *validity, stability, novelty, uniqueness, and diversity*, often relying on inconsistent reference datasets, heuristic filters, or mixed energy estimators. As a result, reported metrics are frequently incomparable across works and may overestimate real-world discovery potential.

We introduce **LeMat-GenBench**, the first unified, open benchmarking framework for **unconditional crystal structure generation**. LeMat-GenBench provides a standardized evaluation pipeline and a comprehensive metric suite designed to assess generative models under realistic discovery conditions.

2. LeMat-GenBench Evaluation Pipeline

Figure 1 illustrates the LeMat-GenBench evaluation pipeline, which transforms raw outputs from generative models into a consistent set of discovery-oriented metrics. The pipeline consists of 3 stages:

1. **Stage 1 - Validity Processing**: Generated structures are first filtered using rigorous physical and chemical validity checks, including charge neutrality, minimum interatomic distance constraints, and basic crystallographic plausibility. This step ensures that downstream metrics are not inflated by unphysical outputs.

2. **Stage 2 - Structural and Energetic Characterization**: Valid structures are evaluated using an ensemble of machine-learned interatomic potentials (MLIPs) [15, 16, 17]. These MLIPs provide estimates of formation energy, relaxation RMSD, and energy above the convex hull (E_{hull}), enabling scalable stability assessment without

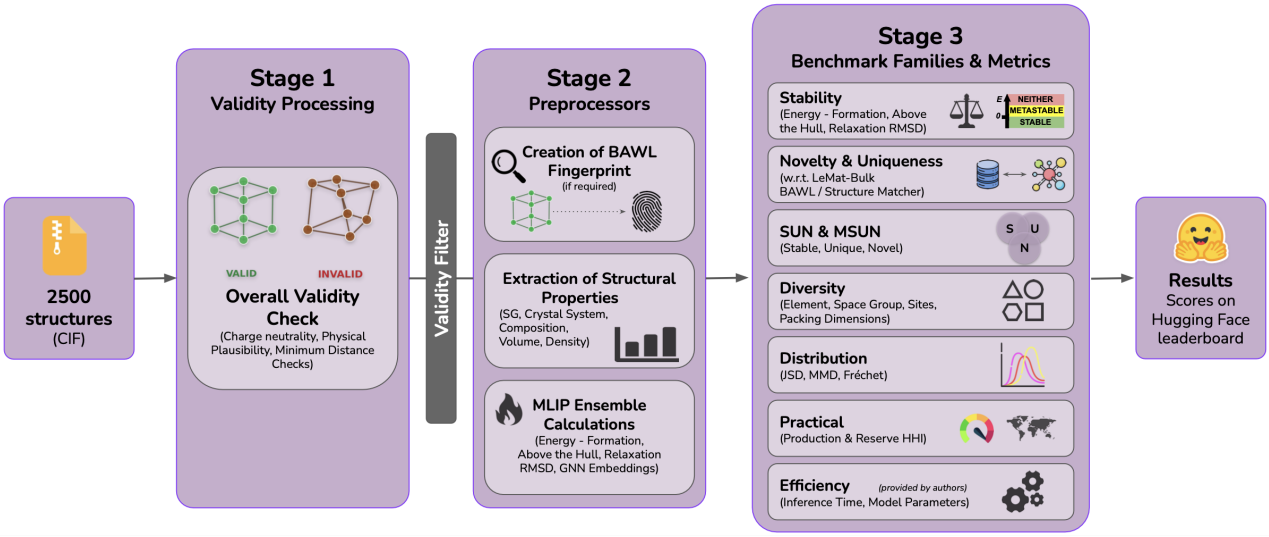

Fig. 1: LeMat-GenBench pipeline from raw generated crystal structures to comprehensive evaluation. The framework applies validity filtering, structural preprocessing, MLIP-based energetic evaluation, and multi-dimensional benchmarking metrics. Results for 15+ models have been added to the leaderboard since release.

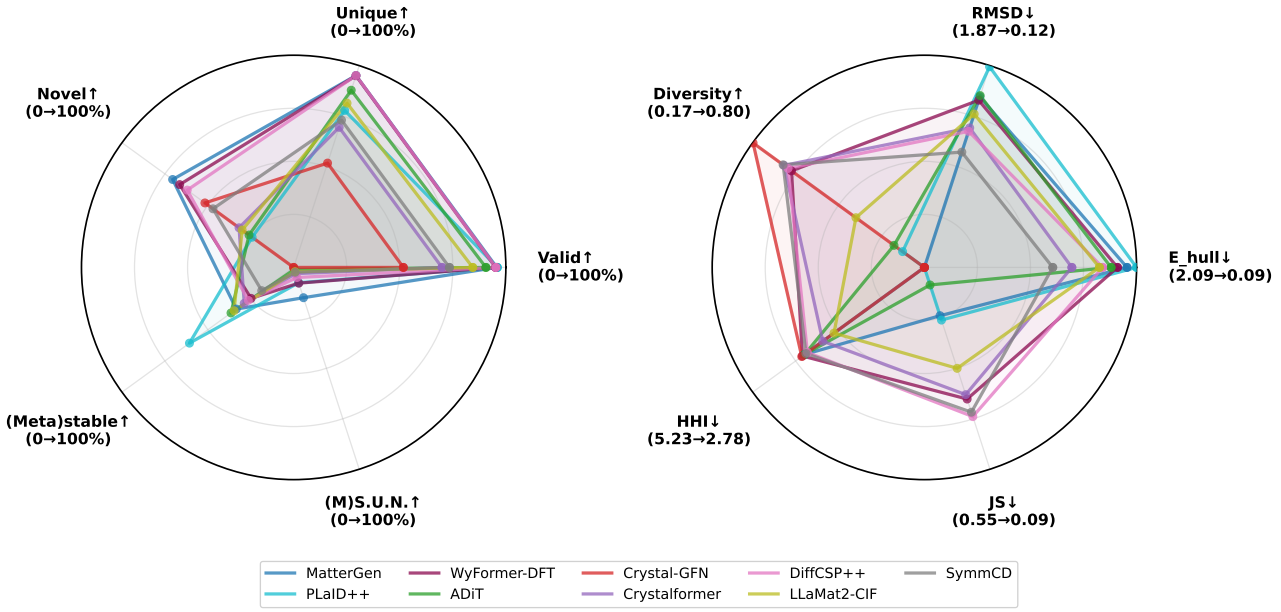

Fig. 2: Spider plots summarizing performance across metric families for representative models.

invoking expensive DFT calculations.

3. **Stage 3 - Benchmark Families and Metrics**: LeMat-GenBench then computes metrics capturing stability, novelty, uniqueness, diversity, distributional alignment, practical synthesizability considerations, and model efficiency.

3. Reference Dataset: LeMat-Bulk

To avoid inflated novelty and stability estimates arising from limited phase coverage, LeMat-GenBench grounds its evaluation in LeMat-Bulk [18], a large-scale dataset comprising approximately 5 million crystal structures aggregated from multiple quantum-chemistry databases.

LeMat-Bulk provides broader compositional and structural coverage than commonly used references such as MP-20 [1], leading to more stringent and realistic assessments of discovery performance. We see that novelty and stability scores can drop by factors of 2–3× [19] when evaluated against LeMat-Bulk instead of smaller references, highlighting the importance of comprehensive phase coverage.

4. Self-Consistent Stability Evaluation

Accurate stability assessment is particularly challenging for generative models. Mixing total energies predicted by Machine Learning Interatomic Potentials (MLIPs) with convex hulls constructed from DFT references introduces systematic inconsistencies, as MLIPs and DFT operate on different energy baselines.

To address this issue, LeMat-GenBench employs a self-consistent MLIP-based convex hull, in which both candidate structures and reference phases are evaluated using the same MLIP ensemble. This approach significantly improves the reliability of E_{hull} estimates and stability classification, particularly near tight thresholds (< 0.01 eV/atom) (Figure A1), where systematic biases can otherwise dominate.

5. Benchmarking Generative Models

We apply LeMat-GenBench to evaluate 12 state-of-the-art crystal generative models [20, 9, 21, 22, 23, 24, 12, 25, 10, 11], spanning GFlowNets, diffusion-based, RL- and LLM-guided approaches.

The results (Figure 2) reveal a consistent trade-off between stability and novelty. Models optimized for thermodynamic plausibility tend to generate conservative candidates close to known phases, while more exploratory models achieve higher diversity and novelty at the cost of stability. No single model dominates across all dimensions, underscoring the need for multi-objective optimization.

6. Conclusions and Outlook

LeMat-GenBench addresses a central bottleneck in generative materials discovery: the absence of shared, realistic, and reproducible evaluation standards. By combining a large-scale reference dataset, self-consistent stability assessment, and a unified metric suite, the benchmark enables fair comparison across diverse generative paradigms.

The current release focuses on unconditional crystal generation, establishing a necessary baseline before meaningful evaluation of conditional or property-guided discovery workflows. Future extensions will incorporate conditional benchmarks, synthesis-aware constraints, and tighter integration with experimental validation pipelines.

By releasing both an open-source evaluation toolkit and a public leaderboard, LeMat-GenBench provides a foundation for community-driven progress toward reliable, discovery-oriented generative models for crystalline materials.

References

[1] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin a. Persson. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013.

[2] J-M Tarascon and Michel Armand. Issues and challenges facing rechargeable lithium batteries. *nature*, 414(6861):359–367, 2001.

[3] 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.

[4] 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.

[5] David S. Sholl and Janice A. Steckel. *Density Functional Theory: A Practical Introduction*. Wiley, March 2009.

[6] 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.

[7] 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.

[8] 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*, 2023.

[9] 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.

[10] 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*, 2025.

[11] 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.

[12] 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.

[13] Cyprien Bone, Matthew Walker, Kuangdai Leng, Luis M Antunes, Ricardo Grau-Crespo, Amil Aligayev, Javier Dominguez, and Keith T Butler. Discovery and recovery of crystalline materials with property-conditioned transformers. *arXiv preprint arXiv:2511.21299*, 2025.

[14] Philipp Höllmer, Thomas Egg, Maya M Martirossyan, Eric Fuemmeler, Zeren Shui, Amit Gupta, Pawan Prakash, Adrian Roitberg, Mingjie Liu, George Karypis, et al. Open materials generation with stochastic interpolants. *arXiv preprint arXiv:2502.02582*, 2025.

[15] 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.

[16] Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M Elena, Dávid P Kovács, Janosh Riebesell, Xavier R Advincula, Mark Asta, Matthew Avaylon, William J Baldwin, et al. A foundation model for atomistic materials chemistry. *The Journal of chemical physics*, 163(18), 2025.

[17] 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.

[18] Martin Siron, Inel Djafar, Ali Ramlaoui, Etienne du Fayette, Amandine Rossello, Edvin Fako, Matthew McDermott, Felix Therrien, Luis Barroso-Luque, Flaviu Cipcigan, et al. Lematbulk: aggregating, and de-duplicating quantum chemistry materials databases. *arXiv preprint arXiv:2511.05178*, 2025.

[19] Siddharth Betala, Samuel P Gleason, Ali Ramlaoui, Andy Xu, Georgia Channing, Daniel Levy, Clémentine Fourrier, Nikita Kazeev, Chaitanya K Joshi, Sékou-Oumar Kaba, et al. Lematgenbench: A unified evaluation framework for crystal generative models. *arXiv preprint arXiv:2512.04562*, 2025.

[20] 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.

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

[22] 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.

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

[24] 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.

[25] 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*, 2025.

Appendix A.

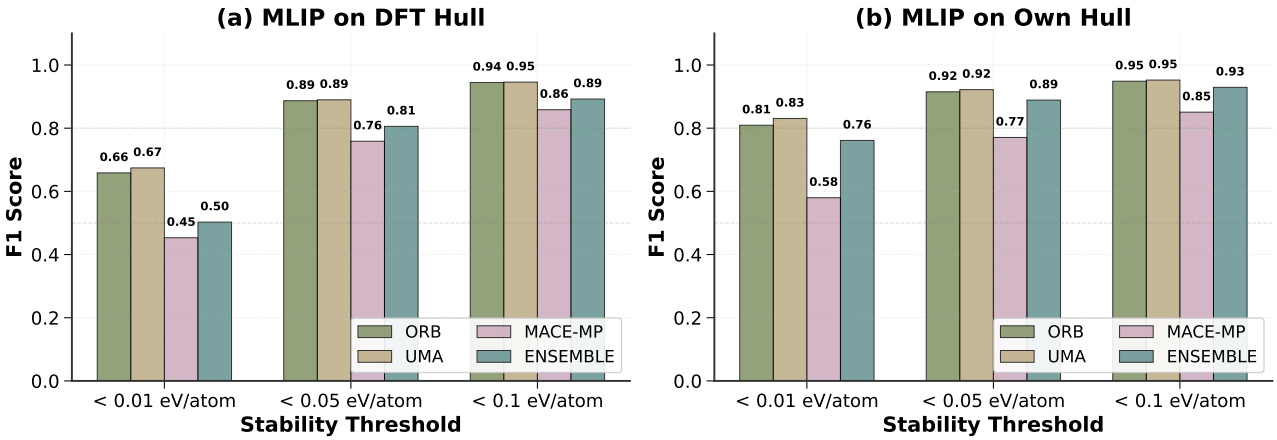

Fig. A1: Comparison of stability prediction using (a) MLIP energies evaluated against a DFT-based hull and (b) a self-consistent MLIP-based hull. The self-consistent approach yields higher F1-scores and lower MAE across stability thresholds.