Continuous Uniqueness and Novelty Metrics for Generative Modeling of Inorganic Crystals

Masahiro Negishi
Imperial College London
m.negishi25@imperial.ac.uk

Hyunsoo Park
Imperial College London
hpark@imperial.ac.uk

Kinga O. Mastej
Imperial College London
k.mastej24@imperial.ac.uk

Aron Walsh
Imperial College London
a.walsh@imperial.ac.uk

Abstract

To address pressing scientific challenges such as climate change, increasingly sophisticated generative artificial intelligence models are being developed that can efficiently sample the large chemical space of possible functional materials. These models can quickly sample new chemical compositions paired with crystal structures. They are typically evaluated using uniqueness and novelty metrics, which depend on a chosen crystal distance function. However, the most prevalent distance function has four limitations: it fails to quantify the degree of similarity between compounds, cannot distinguish compositional difference and structural difference, lacks Lipschitz continuity against shifts in atomic coordinates, and results in a uniqueness metric that is not invariant against the permutation of generated samples. In this work, we propose using two continuous distance functions to evaluate uniqueness and novelty, which theoretically overcome these limitations. Our experiments show that these distances reveal insights missed by traditional distance functions, providing a more reliable basis for evaluating and comparing generative models for inorganic crystals.

1 Introduction

A central challenge in materials science is the efficient design of functional crystals from the vast space of candidate chemical compositions and three-dimensional structures. Traditionally, researchers have relied heavily on chemical knowledge to select promising candidates, followed by experimental or computational validation. However, such approaches are often prohibitively slow in addressing pressing energy and environmental challenges such as climate change. Recently, machine learning (ML) generative models have gained traction for their ability to rapidly generate numerous candidate compounds by learning from large databases of known crystals [1].

Two primary metrics to evaluate generative models for crystals are uniqueness and novelty. Uniqueness measures the diversity within a set of generated samples, while novelty quantifies their dissimilarity from the training data. Since both metrics depend on a defined distance function between crystals, selecting an appropriate function is critical for meaningful assessments. Currently, the most common distance function d_{smat} is based on the "fit" method of the StructureMatcher class in the pymatgen library [2]. d_{smat} returns a Boolean value (`True` if the structures are considered equivalent, `False` otherwise) based on predefined matching thresholds, and therefore fails to quantify the degree of similarity. This is undesirable because the physical properties of materials often vary continuously with gradual changes in crystal structure. Another disadvantage is that d_{smat} does not

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

distinguish between compositional and structural differences. It provides no information as to whether a non-zero distance is due to a difference in composition or structure. In addition, it fails to satisfy Lipschitz continuity against perturbations of atomic positions, which is necessary for robust practical assessments [3]. Furthermore, the uniqueness metric using d_{smat} is not invariant to permutations of the generated samples.

To address these limitations, we propose using two continuous distance functions (compositional and structural), better capturing subtle smooth variations in crystal structures. Section 3 demonstrates how these functions overcome the drawbacks of d_{smat}. Section 4 then experimentally shows that our proposed metrics offer insights that are missed by d_{smat} and other conventional distance functions.

2 Uniqueness and Novelty

Uniqueness A crucial requirement for generative models is the generation of diverse, non-redundant samples. The "uniqueness" of the generated samples quantifies this property. Given a set[1] of generated crystals $X \coloneqq \{x_1, x_2, \ldots, x_n\}$, there are two approaches to measuring uniqueness:

$$\begin{aligned} \text{discrete uniqueness} &\coloneqq \frac{1}{n}\sum_{i=1}^{n} I\left(\bigwedge_{j=1}^{i-1}\left(d_{\mathrm{discrete}}(x_i, x_j) \neq 0\right)\right), \\ \text{continuous uniqueness} &\coloneqq \frac{1}{\binom{n}{2}}\sum_{i=1}^{n}\sum_{j=1}^{i-1} d_{\mathrm{continuous}}(x_i, x_j), \end{aligned} \tag{1}$$

where d_{discrete} is an arbitrary discrete distance function (0 or 1), I is an indicator function that returns 1 if the proposition is true and 0 otherwise, and $d_{\mathrm{continuous}}$ is an arbitrary real-valued distance function [2]. In both cases, a larger value indicates a more diverse set of samples. Many previous studies [4–12] used discrete uniqueness based on the StructureMatcher class of pymatgen.

Novelty Uniqueness alone is insufficient for the comprehensive evaluation of crystal generative models. A model may be capable of generating a diverse set of outputs that are, in fact, only minor variations of structures present in the training set. This limitation necessitates the use of another key metric: "novelty." Novelty measures how different a set of generated samples, X, is from the training data $Y_{\mathrm{train}} \coloneqq \{y_1, y_2, \ldots y_m\}$. Similarly to uniqueness, novelty can be defined in two ways:

$$\begin{aligned} \text{discrete novelty} &\coloneqq \frac{1}{n}\sum_{i=1}^{n} I\left(\bigwedge_{j=1}^{m}\left(d_{\mathrm{discrete}}(x_i, y_j) \neq 0\right)\right), \\ \text{continuous novelty} &\coloneqq \frac{1}{n}\sum_{i=1}^{n} \min_{j=1\sim m} d_{\mathrm{continuous}}(x_i, y_j). \end{aligned} \tag{2}$$

A higher score indicates greater dissimilarity between X and Y_{train}. Again, discrete novelty based on the StructureMatcher class of pymatgen has been the most common choice [4–12].

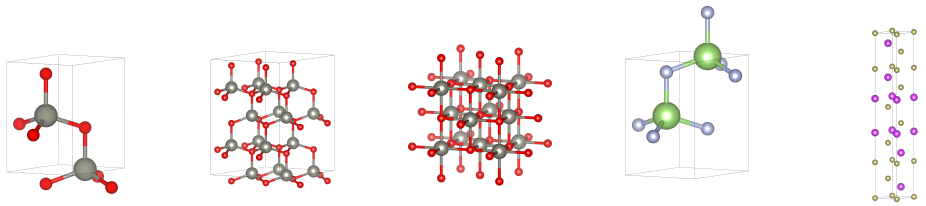
3 Theoretical Analysis of Crystal Distance Functions

In this section, we propose using two continuous distance functions for uniqueness and novelty evaluation. We will show their theoretical advantages over d_{smat} and other distance functions used in prior work. Figure 1 and Table 1 are examples that illustrate how these distances behave. See Appendix A for the detailed definitions and theoretical analyses of each distance.

The most popular distance function d_{smat} operates as follows: if the two given structures have different compositions, d_{smat} returns 1. Otherwise, it reduces both structures to their primitive cells and explores all possible lattice transformations and translations to align one structure onto the other. A match is confirmed (returning a distance of 0) if the maximum distance between corresponding

[1] Mathematically speaking, X is a multiset rather than a mere set because the same crystal can be generated multiple times. However, for simplicity, we refer to it as a set.

[2] To be mathematically precise, d_{discrete} and $d_{\mathrm{continuous}}$ are not distances but pseudometrics. In other words, two different crystals can be considered equivalent. However, for simplicity, we call them distances.



(a) wz-ZnO (b) wz-ZnO* (c) rs-ZnO (d) wz-GaN (e) Bi_2Te_3

Figure 1: Unit cell (and $2 \times 2 \times 2$ supercell*) of example crystal structures used in Table 1.

Table 1: Distances between example pairs of crystals based on composition and structure metrics.

	discrete			continuous	
	d_{smat}	d_{comp}	d_{wyckoff}	d_{magpie}	d_{amd}
wz-ZnO vs wz-ZnO (supercell)	0	0	0	0.000×10^0	4.441×10^{-15}
wz-ZnO vs rs-ZnO	1	0	1	0.000×10^0	1.042×10^0
wz-ZnO vs wz-GaN	1	1	0	6.298×10^2	9.684×10^{-2}
wz-ZnO vs Bi_2Te_3	1	1	1	1.070×10^3	3.240×10^0

atoms, minimized across all potential alignments, falls below a predefined threshold. If no such alignment is found, the function returns a distance of 1.

A key limitation of d_{smat} is that a non-zero distance does not specify whether the difference stems from composition or structure. To enable a more targeted assessment, other distance functions have been introduced. For compositional comparison, d_{comp} returns 0 if and only if two structures share the same composition [13–15]. Similarly, for structural comparison, d_{wyckoff} returns 0 if and only if they share the same space group and Wyckoff letters for the atomic sites in the unit cell [6, 7].

Even though d_{comp} and d_{wyckoff} allow for more targeted evaluations, they are still discrete and cannot measure the degree of similarity. To overcome this limitation, we propose using the following continuous distance functions. For compositional comparison, we suggest using d_{magpie}: the Euclidean distance between Magpie fingerprints consisting of 145 attributes, including stoichiometric attributes and elemental property statistics [16]. Note that Nguyen et al. [17] used d_{magpie} for uniqueness evaluation, but did not apply it to novelty or compare it with other distance functions. For structural comparison, we propose using d_{amd}, the L_∞ distance between Average Minimum Distance (AMD) vectors [18]. Each AMD vector is a structural fingerprint for a crystal, where AMD$[k]$ represents the mean distance from an atom to its kth nearest neighbor, averaged over all atoms in a primitive unit cell. Note that previous studies used d_{amd} to evaluate novelty of a single sample [19], or to identify duplicates in crystal databases or generated samples [20]. However, these studies did not derive a single, continuous uniqueness or novelty score for a set of generated samples to compare different models or distance functions.

The advantages of d_{magpie} and d_{amd} extend beyond their real-valued nature. Both functions satisfy two crucial properties for robust evaluation, as outlined by [3].

- Isometry invariance: for any two isometric crystals, $x \cong x'$, the distance $d(x, x') = 0$.
- Lipschitz continuity: there exists a positive constant C s.t. if x' is obtained from x by shifting each atom by a distance of at most ε, then $d(x, x') \leq C\varepsilon$.

Isometry invariance ensures that identical crystals are not misclassified as different. Meanwhile, Lipschitz continuity guarantees that small perturbations in atomic positions lead to only small changes in the distance. It is essential for practical robustness because it accounts for factors such as atomic vibrations, experimental measurement errors, and fluctuations in the generative model's outputs. d_{smat} does not satisfy Lipschitz continuity because it compares the primitive unit cells of crystals, which change discontinuously as the atomic positions change. On the other hand, d_{comp} and d_{magpie} trivially meet both requirements since they are compositional. d_{amd} also satisfies both properties thanks to its invariance with respect to the choice of the unit cell [18]. This gives d_{amd} a clear

advantage over d_{wyckoff}, which satisfies neither property. Wyckoff letters are defined relative to the chosen origin and setting of the space group, so they may change if the origin of the unit cell is shifted or a different setting is used[3]. Note that the Euclidean distance between structural CrystalNN fingerprints [21], which have been used for uniqueness and novelty evaluation [14, 17], also fails to satisfy Lipschitz continuity. The same is true for another popular continuous structural distance: the regularized-entropy match distance using structural Smooth Overlap of Atomic Positions (SOAPs) descriptors [22]. This motivates us to choose d_{amd} as a continuous structural distance.

Before presenting our experimental results, we highlight one final drawback of the prevalent d_{smat}: its uniqueness score is not invariant against the permutation of generated samples. Ideally, the evaluation of a set of samples should not be affected by their generation order; whether structure x is generated before structure x' should be irrelevant to the final score. However, the uniqueness value calculated with d_{smat} violates this fundamental principle. This means that different uniqueness scores can be obtained for the same set of crystals simply by shuffling their order. To illustrate this, consider a set of three generated samples x, x', and x'', where $d_{\mathrm{smat}}(x, x') = d_{\mathrm{smat}}(x, x'') = 0$, but a cumulative difference leads to $d_{\mathrm{smat}}(x', x'') = 1$. If the generation order is $x \to x' \to x''$, the calculated uniqueness is $\frac{1}{3}$, as only x is considered unique. However, if they are sampled as $x' \to x'' \to x$, both x' and x'' are regarded as unique, giving a uniqueness value of $\frac{2}{3}$. This lack of permutation invariance stems from the fact that d_{smat} is not a true mathematical pseudometric. Specifically, it violates the triangle inequality. None of the other crystal distance functions introduced in this paper have this flaw.

4 Experiments

We evaluated six generative models (CDVAE [23], DiffCSP [24], DiffCSP++ [25], MatterGen [4], Chemeleon-DNG [26], and ADiT [10]), each trained on the MP20 dataset [23]. These are each used to generate 10k structures based on unguided (de novo) sampling of crystal space. We then computed the uniqueness and novelty metrics on the samples using different distance functions. We also performed evaluations on a filtered subset of thermodynamically metastable structures with the energy above the convex hull $E_{\mathrm{hull}} \leq 0.1$ [eV/atom] to specifically assess the ability of each model to generate stable and unique/novel samples. Note that n in the denominators of equations 1 and 2 was fixed to 10k in all calculations. See Appendix B.1 for more detailed settings. The codes for the experiments and the Python package are available online (https://github.com/WMD-group/xtalmet).

A key observation from the results presented in Table 2 is the strong correlation between the metrics based on d_{smat} and d_{comp}. In other words, models performing well on d_{smat}-based metrics also tend to perform well on d_{comp}-based metrics, and vice versa. This implies that d_{smat} returns 1 primarily due to compositional differences rather than structural ones. Another significant observation is that uniqueness based on continuous distances reveals weaknesses that conventional discrete metrics cannot. For example, without screening, CDVAE reaches the highest uniqueness in d_{comp} while performing worst on the more refined d_{magpie}-based compositional uniqueness. This suggests that while CDVAE's outputs have few exact compositional duplicates, their overall distribution is highly concentrated rather than dispersed, which hinders diverse candidate generation. Uniqueness based on discrete d_{comp} undesirably highly evaluates such non-scattered samples. A similar trend appears in structural uniqueness, where DiffCSP++ is the top-performing model when evaluated with the discrete d_{wyckoff}, yet it scores poorly with the continuous d_{amd}. A third finding is that uniqueness and novelty evaluations should be combined with stability. CDVAE provides a clear example: its performance drops dramatically after we filter for thermodynamic stability. This is because only about 3% of CDVAE's samples pass the screening, which is far lower than the other models (see Table 4 in Appendix B.2). Essentially, CDVAE achieves high creativity scores by simply producing a variety of physically unrealistic structures. Once stability is taken into account, Chemeleon-DNG becomes the top-performing model across all uniqueness metrics, while MatterGen excels in most novelty metrics. In Appendix B.2, we present an analysis of computational time, as well as a discussion on how to evaluate uniqueness and novelty simultaneously. We also experimentally demonstrate that d_{smat}-based uniqueness score is not invariant to the generation order in practice, as we show theoretically at the end of Section 3.

[3]Computing d_{wyckoff} on Niggli's reduced cells makes it isometry-invariant. However, previous studies that used d_{wyckoff} to evaluate their generative models did not perform this reduction [6, 7]. Therefore, we will not perform it in this paper. Note that d_{wyckoff} is not Lipschitz continuous even after Niggli reduction.

Table 2: Uniqueness (U) and Novelty (N) results. The shaded rows represent the results after screening based on the energy above the convex hull E_{hull}, and the best model is underlined in each row. Note that, by definition, the uniqueness and novelty defined with d_{magpie} or d_{amd} can exceed one. Results for d_{magpie} are scaled to 1/1000 for better visibility.

		CDVAE	DiffCSP	DiffCSP++	MatterGen	Chemeleon-DNG	ADiT
U	d_{smat}	0.9945	0.9771	0.9806	0.9838	0.9791	0.8841
	d_{comp}	0.9717	0.9461	0.9522	0.9523	0.9373	0.7743
	d_{wyckoff}	0.0045	0.0324	0.1910	0.0470	0.0637	0.0038
	d_{magpie}	1.7951	1.9816	2.0704	2.0890	2.0839	2.0735
	d_{amd}	1.2069	1.5912	1.3768	1.4145	2.6791	1.2725
	d_{smat}	0.0346	0.2885	0.2723	0.3517	0.3747	0.3164
	d_{comp}	0.0341	0.2775	0.2646	0.3372	0.3599	0.3089
	d_{wyckoff}	0.0021	0.0209	0.0423	0.0371	0.0439	0.0033
	d_{magpie}	0.0020	0.1773	0.1597	0.2530	0.2975	0.2695
	d_{amd}	0.0016	0.1376	0.1175	0.2010	0.2268	0.1832
N	d_{smat}	0.9892	0.9050	0.8925	0.9158	0.8630	0.4582
	d_{comp}	0.9221	0.8207	0.8121	0.8156	0.7497	0.2352
	d_{wyckoff}	0.0352	0.0812	0.0098	0.0967	0.1134	0.0425
	d_{magpie}	0.1063	0.0848	0.0883	0.0773	0.0678	0.0155
	d_{amd}	0.1973	0.2182	0.1291	0.1423	0.6986	0.1359
	d_{smat}	0.0319	0.2190	0.1890	0.2845	0.2621	0.0722
	d_{comp}	0.0295	0.1816	0.1619	0.2294	0.2158	0.0604
	d_{wyckoff}	0.0012	0.0368	0.0013	0.0554	0.0512	0.0323
	d_{magpie}	0.0031	0.0160	0.0133	0.0188	0.0168	0.0038
	d_{amd}	0.0075	0.0222	0.0168	0.0396	0.0258	0.0431

5 Conclusion

We have argued that the two continuous distance functions, d_{magpie} and d_{amd}, are superior to the widely used d_{smat} for evaluating uniqueness and novelty. This is because they enable non-binary, and compositionally or structurally targeted assessments. They are also theoretically robust to slight changes in atomic positions, and their uniqueness scores are invariant to the generation order. Our experiments have shown that d_{smat} primarily measures compositional difference, and continuous distances can better distinguish generative model performance for inorganic crystals.

A key limitation of continuous metrics is that, unlike their binary counterparts, they do not provide a straightforward way to screen individual generated materials. Developing new screening techniques based on continuous distances is an important future direction. Additionally, while we adopted d_{magpie}, other continuous compositional distances, such as the learned embeddings of foundation machine learning force fields, offer the same advantages as d_{magpie}. Exploring these options is also a promising next step. Finally, we plan to apply our metrics, particularly the structural d_{amd}, to evaluate for crystal structure prediction tasks.

Acknowledgments and Disclosure of Funding

We thank the EPSRC for support via the AI for Chemistry: AIchemy Hub (EPSRC grant EP/Y028775/1 and EP/Y028759/1).

References

[1] Hyunsoo Park, Zhenzhu Li, and Aron Walsh. Has generative artificial intelligence solved inverse materials design? *Matter*, 7(7):2355–2367, 2024.

[2] 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.

[3] Daniel Widdowson and Vitaliy Kurlin. Resolving the data ambiguity for periodic crystals. In *Advances in Neural Information Processing Systems*, 2022.

[4] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Chunlei Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, 2025.

[5] Benjamin Kurt Miller, Ricky TQ Chen, Anuroop Sriram, and Brandon M Wood. Flowmm: Generating materials with riemannian flow matching. In *Forty-first International Conference on Machine Learning*, 2024.

[6] 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.

[7] 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.

[8] Anuroop Sriram, Benjamin Miller, Ricky TQ Chen, and Brandon Wood. Flowllm: Flow matching for material generation with large language models as base distributions. *Advances in Neural Information Processing Systems*, 37:46025–46046, 2024.

[9] Pierre-Paul De Breuck, Hashim A Piracha, Gian-Marco Rignanese, and Miguel AL Marques. A generative material transformer using wyckoff representation. *arXiv preprint arXiv:2501.16051*, 2025.

[10] 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 Machine Learning*, 2025.

[11] François R J Cornet, Federico Bergamin, Arghya Bhowmik, Juan Maria Garcia-Lastra, Jes Frellsen, and Mikkel N. Schmidt. Kinetic langevin diffusion for crystalline materials generation. In *Forty-second International Conference on Machine Learning*, 2025.

[12] Philipp Höllmer, Thomas Egg, Maya Martirossyan, Eric Fuemmeler, Zeren Shui, Amit Gupta, Pawan Prakash, Adrian Roitberg, Mingjie Liu, George Karypis, Mark Transtrum, Richard Hennig, Ellad B. Tadmor, and Stefano Martiniani. Open materials generation with stochastic interpolants. In *Forty-second International Conference on Machine Learning*, 2025.

[13] 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.

[14] Zekun Ren, Siyu Isaac Parker Tian, Juhwan Noh, Felipe Oviedo, Guangzong Xing, Jiali Li, Qiaohao Liang, Ruiming Zhu, Armin G Aberle, Shijing Sun, et al. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter*, 5(1):314–335, 2022.

[15] 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.

[16] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2(1):1–7, 2016.

[17] Tri Minh Nguyen, Sherif Abdulkader Tawfik, Truyen Tran, Sunil Gupta, Santu Rana, and Svetha Venkatesh. Hierarchical gflownet for crystal structure generation. In *AI for Accelerated Materials Design-NeurIPS 2023 Workshop*, 2023.

[18] Daniel Widdowson, Marco M Mosca, Angeles Pulido, Vitaliy Kurlin, and Andrew I Cooper. Average minimum distances of periodic point sets - foundational invariants for mapping periodic crystals. *MATCH Communications in Mathematical and in Computer Chemistry*, 87(3):529–559, 2022.

[19] Daniel Widdowson and Vitaliy Kurlin. Geographic-style maps with a local novelty distance help navigate in the materials space. *Scientific reports*, 15(1):27588, 2025.

[20] Olga D. Anosova, Daniel E. Widdowson, and Vitaliy A. Kurlin. Recognition of near-duplicate periodic patterns by continuous metrics with approximation guarantees. *Pattern Recognition*, 171:112108, 2026.

[21] Nils ER Zimmermann and Anubhav Jain. Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity. *RSC advances*, 10(10):6063–6081, 2020.

[22] Sandip De, Albert P Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics*, 18(20): 13754–13769, 2016.

[23] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi S. Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. In *International Conference on Learning Representations*, 2022.

[24] 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.

[25] Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation. In *The Twelfth International Conference on Learning Representations*, 2024.

[26] Hyunsoo Park, Anthony Onwuli, and Aron Walsh. Exploration of crystal chemical space using text-guided generative artificial intelligence. *Nature Communications*, 16(1):1–14, 2025.

[27] Donald E Sands. *Introduction to crystallography*. Courier Corporation, 1993.

[28] Ilyes Batatia, David Peter Kovacs, Gregor N. C. Simm, Christoph Ortner, and Gabor Csanyi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems*, 2022.

[29] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

Table 3: Properties of each crystal distance function.

	discrete			continuous	
	d_{smat}	d_{comp}	d_{wyckoff}	d_{magpie}	d_{amd}
Compositional	✓	✓		✓	
Structural	✓		✓		✓
Isometry invariance	✓	✓		✓	✓
Lipschitz continuity		✓		✓	✓

A Crystal Distance Functions

This section provides precise definitions and properties of the crystal distance functions used in the main text. Table 3 summarizes these properties.

A.1 Definitions

$\boldsymbol{d_{\mathrm{smat}}}$ is the most widely used discrete distance, derived from the "fit" method of the StructureMatcher class in the pymatgen library [2]. The algorithm proceeds as follows:

1. The function first checks for compositional identity. If the two input crystals have different compositions, it returns a distance of 1.
2. If the compositions match, the algorithm reduces the input structures to their primitive cells. Then, it systematically searches through all possible lattice transformations and translations to align one structure with the other. A match is declared (distance = 0) if an alignment is found where all corresponding atoms are of the same chemical element and the distance between each matched atom pair is below a predefined threshold. Otherwise, the function returns a distance of 1.

Essentially, the first step checks for compositional differences, while the second checks for structural differences. Consequently, when applied to the examples in Table 1, this function returns a distance of 1 for every pair except for the first one (wz-ZnO and its supercell), which is compositionally and structurally identical. Note that d_{smat} is mathematically not a pseudometric because it can violate the triangle inequality. For example, consider three crystals, x, x', and x''. It is possible for the difference between x and x' to be below the matching threshold, and the same for x' and x'', resulting in $d_{\mathrm{smat}}(x, x') = 0$ and $d_{\mathrm{smat}}(x', x'') = 0$. However, the cumulative difference could mean that $d_{\mathrm{smat}}(x, x'') = 1$, which violates the inequality ($0 + 0 \ngeq 1$). Nevertheless, for simplicity, we refer to d_{smat} as a "distance" throughout this paper, as we do with other pseudometric functions introduced below.

$\boldsymbol{d_{\mathrm{comp}}}$ checks whether the compositions of the input structures are identical:

$$d_{\mathrm{comp}}(x_1, x_2) \coloneqq \begin{cases} 0 & \text{if the compositions are the same,} \\ 1 & \text{otherwise.} \end{cases}$$

As expected, it returns 0 for the first two pairs in Table 1 because each pair shares the same composition, ZnO.

$\boldsymbol{d_{\mathrm{wyckoff}}}$ is a discrete structural distance between crystals. It is defined based on the concepts of space groups and Wyckoff positions. We offer a brief overview of these concepts below. A more detailed and formal discussion can be found in standard crystallography literature, such as [27].

A space group provides a full description of a crystal's symmetry. It encompasses the set of all possible symmetry operations (such as translations, rotations, and reflections) that leave the crystal unchanged. There are 230 unique space groups in total, each representing a distinct set of these operations. Each crystal can be classified into one of these groups. For example, wz-ZnO in Figure 1(a) has symmetry operations such as a six-fold screw axis along the c-axis, and is classified as space group 186.

Once the space group is determined, not every atomic position is unique. Applying the symmetry operations of the space group to an atom at a specific coordinate will generate a set of symmetrically

equivalent positions. A Wyckoff position is the label assigned to such a set. In other words, it represents a specific type of location (e.g., a point, line, or plane) within the unit cell. Once an atom is placed at one of these locations, all of its equivalent positions are automatically determined by the space group's symmetry. Each space group has a finite list of possible Wyckoff positions, denoted by letters (a, b, c, etc.), which correspond to sites with different symmetries. Consider the example of wz-ZnO in Figure 1(a). The unit cell of wz-ZnO contains two Zn atoms and two O atoms. Both atoms independently occupy the Wyckoff position denoted by "b," whose representative coordinate is $(\frac{1}{3}, \frac{2}{3}, z)$. Applying the symmetry operations of space group 186 to this single coordinate generates the second, equivalent position at $(\frac{2}{3}, \frac{1}{3}, z+\frac{1}{2})$. Therefore, it is unnecessary to specify the coordinates for all four atoms. Defining one representative atom for each set of equivalent positions is sufficient. The free parameter z is approximately 0 for Zn and 0.38 for O.

In summary, a crystal can be fully described by its space group, lattice parameters (the lengths and angles of the three lattice vectors), the occupied Wyckoff positions, and the chemical elements at those positions. For instance, wz-ZnO in Figure 1(a) can be represented as follows:

$$\begin{aligned}&\text{space group} : 186\\&\text{lattice parameters} : (3.24\text{Å},\ 3.24\text{Å},\ 5.22\text{Å},\ 90^\circ,\ 90^\circ,\ 120^\circ)\\&\text{Wyckoff positions and elements} : \text{Zn @ } b\ (z=0), \text{O @ } b\ (z=0.38).\end{aligned}$$

d_{wyckoff} checks whether two crystals share the same space group and the same multiset of occupied Wyckoff letters:

$$d_{\text{wyckoff}}(x_1, x_2) \coloneqq \begin{cases} 0 & \text{if they have the same space group and the Wyckoff letters,} \\ 1 & \text{otherwise.} \end{cases}$$

In the example of wz-ZnO, it only uses the information of its space group (186), and the multiset of Wyckoff letters, $\{\{b, b\}\}$. It discards the compositional information and only considers (a part of) structural information. For the examples in Table 1, both the first and third pairs have the same space group and Wyckoff letters, so their d_{wyckoff} distance is 0. Note that the exact coordinates of atoms are different for the third pair.

$\boldsymbol{d_{\text{magpie}}}$ allows for continuous measurement of compositional distance. It is defined as the L_2 distance between fixed-length vectors, called Magpie fingerprints [16], i.e.,

$$d_{\text{magpie}}(x_1, x_2) \coloneqq \|M(x_1) - M(x_2)\|_2,$$

where M is a function that converts a crystal to its Magpie fingerprint. Magpie fingerprints consist of 145 attributes that fall into four distinct categories: 1) stoichiometric attributes, which depend only on the elemental fractions; 2) elemental property statistics, such as the mean, range, and mode of 22 elemental properties; 3) electronic structure attributes, which represents the average valence electron contributions from the s, p, d, and f shells; and 4) ionic compound attributes, which include a feature for the possibility of forming an ionic compound and two measures of fractional ionic character derived from electronegativity. Similar to d_{comp}, d_{magpie} also returns a distance of 0 for the first two pairs in Table 1, because the Magpie fingerprint depends solely on composition. However, the advantage of a continuous metric becomes clear with the third and fourth pairs. For these examples, d_{magpie} provides a more nuanced comparison of compositional similarity, whereas the binary d_{comp} can only indicate that they are different.

$\boldsymbol{d_{\text{amd}}}$ is a continuous structural distance defined as the L_∞ distance between structural fingerprints, called Average Minimum Distance (AMD) vectors [18]:

$$d_{\text{amd}}(x_1, x_2) \coloneqq \|\text{AMD}(x_1) - \text{AMD}(x_2)\|_\infty,$$

where $\text{AMD}(x)$ is the AMD vector of x. The kth element of an AMD vector stands for the average distance from an atom to its kth nearest neighbor, with the average taken over all atoms in the structure. In theory, d_{amd} should be zero for the first pair in Table 1, since the two structures are identical. In practice, however, it returns a tiny non-zero value. This discrepancy is due to the finite precision of Python's floating-point calculations. For the remaining three pairs, d_{amd} returns non-zero values as expected. Of these three pairs, (wz-ZnO, wz-GaN) exhibits the smallest distance. This result makes sense, as this pair is the only one that shares the same space group and Wyckoff letters. Compared with d_{wyckoff}, d_{amd} does not discard the information about exact coordinates and serves as a more detailed indicator of structural similarity.

A.2 Isometry Invariance and Lipschitz Continuity

Here, we take a closer look at the two important properties of crystal distance functions introduced in Section 3. We also analyze whether or not each distance function satisfies these properties.

Isometry Invariance A robust evaluation of uniqueness and novelty requires that a distance function correctly identifies identical crystals, even if they are translated or rotated differently. Without this capability, a single crystal could be counted multiple times when measuring uniqueness, or a generated crystal might be incorrectly flagged as novel when it is actually just a different orientation of a structure already in the training set. This fundamental requirement is called "isometry invariance", and formally defined as follows [3]:

For any two isometric crystals, $x \cong x'$, the distance $d(x, x') = 0$.

The most widely used d_{smat} satisfies isometry invariance because it searches over all possible lattice transformations and translations to align input structures. The compositional distances, d_{comp} and d_{magpie}, also trivially satisfy this property because a crystal's composition is unaffected by isometric transformations. d_{amd} is also isometry-invariant since interatomic distances do not change with the crystal's overall orientation. The one that fails is d_{wyckoff}: Wyckoff letters depend on the choice of the unit cell's origin, so it is not isometry-invariant. You can address this issue by computing d_{wyckoff} on Niggli's reduced cells. However, since previous studies that used d_{wyckoff} to evaluate their generative models did not perform the reduction [6, 7], we do not use it in this paper.

Lipschitz Continuity For a robust evaluation, it is also desirable for a distance function to not be overly sensitive to small perturbations in atomic positions. Many factors can lead to such perturbations in practice, including atomic vibrations, experimental measurement errors, and fluctuations in the generative model's outputs. Widdowson and Kurlin [3] mathematically defined this requirement using the concept of Lipschitz continuity:

There exists a positive constant C s.t. if x' is obtained from x by shifting each atom by a distance of at most ε, then $d(x, x') \leq C\varepsilon$.

The compositional distances, d_{comp} and d_{magpie}, trivially satisfy Lipschitz continuity. Since they are unaffected by atomic positions, they satisfy the condition for any $C > 0$. d_{amd} is also Lipschitz continuous because the underlying interatomic distances change smoothly with atomic perturbations. In contrast, the discrete structural distances do not satisfy this property. d_{smat} fails since the distance suddenly changes from 0 to 1 when the primitive unit cell changes due to atomic perturbations. Similarly, d_{wyckoff} is not Lipschitz continuous because small atomic shifts can alter the calculated space group, which consequently changes the discrete Wyckoff letters.

A.3 Invariance of Uniqueness Against Permutation of Generated Samples

In the final paragraph of Section 3, we argued that d_{smat}-based uniqueness is not invariant to the generation order because d_{smat} violates the triangle inequality. Here, we provide a proof that discrete uniqueness based on a true mathematical pseudometric always satisfies the invariance. Note that continuous uniqueness is always independent of the generation order because it is simply the average distance between generated samples.

Theorem 1 (Discrete Uniqueness with Mathematical Pseudometric Is Permutation Invariant)**.** *Assume d_{discrete} is a true mathematical pseudometric, i.e., satisfying the following properties for any x, x', and x'':*

$$
\begin{aligned}
&d_{\text{discrete}}(x, x) = 0, \\
&d_{\text{discrete}}(x, x') = d_{\text{discrete}}(x', x), \\
&d_{\text{discrete}}(x, x'') \leq d_{\text{discrete}}(x, x') + d_{\text{discrete}}(x', x'').
\end{aligned}
$$

Then, the discrete uniqueness in Equation 1 defined by this d_{discrete} is invariant to the permutation of generated samples.

Proof. We show that a set of generated samples $\{x_1, x_2, \ldots, x_n\}$ can be decomposed into disjoint sets $S_1, S_2, \ldots, S_m (m \leq n)$ that satisfies the following properties:

- The distance between any two samples from the same set is zero.
- The distance between any two samples from different sets is one.

If this holds, then the uniqueness score is always $\frac{m}{n}$ regardless of the generation order, because exactly one sample from each S_i is considered unique. We prove the existence of such disjoint sets by explicitly constructing them. First, we assign x_1 to S_1. Then, we iterate the following procedure for $k = 2 \sim n$: if $d_{\text{discrete}}(x_i, x_k) = 0$ for some $i < k$, x_k is assigned to the same set as x_i. Otherwise, x_k is assigned to a new set. Throughout this process, the set of disjoint sets always satisfies the two requirements. When $n = 1$, $S_1 = x_1$ fulfills them trivially. Suppose the two properties hold for $n = k(1 \leq k \leq n-1)$, and consider assigning x_{k+1}.

Case 1 If x_{k+1} is assigned to an existing set, then that set is uniquely determined. Otherwise, there exist x_i and x_j $(i < j < k)$ from different sets such that:

$$d_{\text{discrete}}(x_i, x_j) \leq d_{\text{discrete}}(x_i, x_k) + d_{\text{discrete}}(x_k, x_j) = 0 + 0 = 0,$$

which contradicts the assumption that the distance between any two samples from different sets is one. To prove that the two requirements hold, it is enough to show that x_{k+1} is at a distance of 0 from any sample within the same set and at a distance of 1 from any sample in a different set. In the set x_{k+1} belongs to, there exists $x_i (i \leq k)$ such that $d_{\text{discrete}}(x_{k+1}, x_i) = 0$, since x_{k+1} is assigned to that set. Then, for any $x_j (j \leq k)$ in the same set,

$$d_{\text{discrete}}(x_{k+1}, x_j) \leq d_{\text{discrete}}(x_{k+1}, x_i) + d_{\text{discrete}}(x_i, x_j) = 0 + 0 = 0.$$

Therefore, the distance between x_{k+1} and any other sample in the same set is 0. Next, assume that there is a sample $x_i (i \leq k)$ from a different set of x_{k+1} that satisfies $d_{\text{discrete}}(x_{k+1}, x_i) = 0$. Then, for any $x_j (j \leq k)$ in the same set as x_{k+1},

$$d_{\text{discrete}}(x_i, x_j) \leq d_{\text{discrete}}(x_i, x_{k+1}) + d_{\text{discrete}}(x_{k+1}, x_j) = 0 + 0 = 0.$$

This contradicts the fact that the distance between any two samples from different sets is one. Thus, x_{k+1} and any sample from a different set have a distance of 1.

Case 2 If x_{k+1} is assigned to a new set, then the two properties hold trivially. □

B Supplementary Materials for Experiments

B.1 Experimental Settings

We evaluated 10k structures from each of six different generative models (CDVAE [23], DiffCSP [24], DiffCSP++ [25], MatterGen [4], Chemeleon-DNG [26], and ADiT [10]) trained on the MP20 dataset. For CDVAE, DiffCSP, DiffCSP++, and MatterGen, we generated the structures ourselves using their public code and checkpoints. For Chemeleon-DNG and ADiT, we utilized the publicly available sets of 10k structures released by the authors.

We computed uniqueness and novelty metrics for two distinct sample sets: 1) the full set of 10k generated structures, and 2) a filtered subset containing only thermodynamically metastable structures. To create the second subset, we first predicted the formation energy of each candidate using the MACE-MPA-0 machine learning potential [28]. We then computed the energy above the convex hull (E_{hull}) relative to the Materials Project database [29]. Any structure with an E_{hull} greater than 0.1 [eV/atom] was considered unstable and removed from the subset.

B.2 Additional Experimental Results

Here, we discuss additional experimental results.

As shown in Table 2 in the main text, the performance of CDVAE declines more sharply than other models after filtering for thermodynamic stability. The reason for this is evident in Table 4, which shows that samples from CDVAE have a much lower stability rate. Specifically, CDVAE samples are approximately 10 times less likely to pass the screening than samples from other models. This indicates that CDVAE has a strong tendency to generate physically implausible structures.

Next, we compare the computational cost of each distance function (Table 5). All calculations were performed on 10k samples generated from CDVAE trained on the MP20 dataset. The results

Table 4: Ratio of samples that are classified as thermodynamically metastable, i.e., $E_{\mathrm{hull}} \leq 0.1$[eV/atom].

CDVAE	DiffCSP	DiffCSP++	MatterGen	Chemeleon-DNG	ADiT
0.0348	0.3087	0.2871	0.3646	0.3909	0.3708

Table 5: Computation time [s] required to calculate the uniqueness and novelty metrics on all the 10k samples generated from CDVAE trained on the MP20 dataset. A single CPU (AMD EPYC 7742) with 32 GB of RAM was used.

	discrete			continuous	
	d_{smat}	d_{comp}	d_{wyckoff}	d_{magpie}	d_{amd}
Uniqueness	1.682×10^4	8.487×10^0	2.918×10^1	6.096×10^2	4.343×10^0
Novelty	4.438×10^4	2.394×10^1	3.968×10^2	2.140×10^3	3.970×10^1

clearly show that evaluations with the commonly used d_{smat} are significantly slower than those with other distances. In fact, our proposed d_{magpie} and d_{amd} are about 10 and 1000 times faster, respectively. This performance gap stems from the fact that distances other than d_{smat} allow for a one-time pre-computation of embeddings. For uniqueness, this involves calculating an embedding (e.g., a Magpie fingerprint or an AMD vector) for each of the n samples. After this initial step, the $\binom{n}{2}$ pairwise distances are calculated between these embeddings, which is computationally trivial (e.g., an L_2 or L_∞ distance). In contrast, d_{smat} does not have a pre-computation step. Therefore, it must perform a computationally intensive comparison for every one of the $\binom{n}{2}$ pairs, resulting in a much higher overall cost. A similar logic applies to novelty calculations. The low computational cost of d_{magpie} and d_{amd} is another advantage compared to d_{smat}.

Next, we discuss how to evaluate uniqueness and novelty at the same time. Until now, we have assessed them separately. This approach has the advantage of providing detailed insights about models, but there exists a certain disadvantage: it makes it difficult to determine an overall top-performing model. To address this, previous studies [4–12] have adopted the S.U.N. metric, which provides a single, unified score. It is defined as the percentage of Stable, Unique, and Novel samples among all generated samples. Stability is determined by E_{hull}, while uniqueness and novelty are based on the discrete d_{smat} distance. However, the S.U.N. metric does not readily extend to continuous distances, such as d_{magpie} and d_{amd}. This is because the S.U.N. metric counts individual "good" samples, but continuous uniqueness does not rate samples one by one. It evaluates a set of generated samples as a whole, instead of assigning a "unique" status to each sample. To simultaneously evaluate uniqueness and novelty with any distance function, we propose using the concept of Pareto optimality. Suppose we want to compare K different models $\{m_k\}_{k=1}^{K}$. Let $U(m_k)$ and $N(m_k)$ represent the uniqueness and novelty of model m_k. These values can be computed using any discrete or continuous distance function. We then consider a model m_k as one of the best-performing models if it is Pareto optimal, meaning that no other model in the set is superior in both uniqueness and novelty. Formally, a model m_k is Pareto optimal if there is no $l(\neq k) \in [1, K]$ such that:

$$\big(U(m_l) \geq U(m_k)\big) \wedge \big(N(m_l) \geq N(m_k)\big) \wedge \Big(\big(U(m_l) > U(m_k)\big) \vee \big(N(m_l) > N(m_k)\big)\Big)$$

While this approach may not identify the single best model, it enables the simultaneous evaluation of uniqueness and novelty with any distance function. If we compute uniqueness and novelty after filtering out thermodynamically unstable materials, as the S.U.N. metric does, then stability can also be reflected. Figure 2 shows the Pareto-optimal models identified using different distance functions. Notably, MatterGen and Chemeleon-DNG always lie on the Pareto frontier regardless of the chosen distance function. Another interesting observation is that the plot for d_{smat} is much closer to that of d_{comp} than d_{wyckoff}. This suggests that, although d_{smat} is defined to capture both compositional and structural differences, it behaves more like a compositional match check in practice.

Finally, we investigate if the lack of permutation invariance in d_{smat}-based uniqueness, which we theoretically proved at the end of Section 3, is a practical concern. For each set of 10k samples generated from the six different models, we created five different permutations by shuffling them with different random seeds. We then calculated the uniqueness score using d_{smat} for each shuffled set.

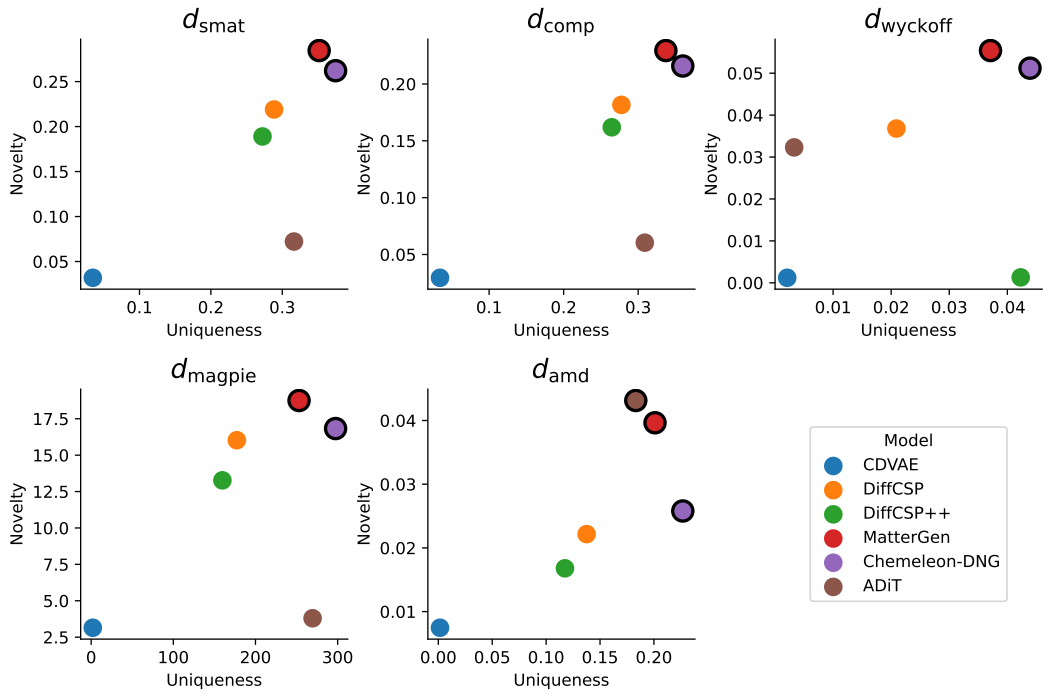


Figure 2: 2D plots showing the uniqueness and novelty of each model, which were computed after screening for thermodynamic stability. Different plots correspond to different distance functions. Pareto-optimal models are highlighted with a black circle.

Table 6: d_{smat}-based uniqueness score calculated on a set of 10k generated samples that were shuffled with different seeds. The final two rows show the average and standard deviation across five seeds.

	CDVAE	DiffCSP	DiffCSP++	MatterGen	Chemeleon-DNG	ADiT
seed = 0	0.9942	0.9770	0.9809	0.9839	0.9792	0.8839
seed = 1	0.9945	0.9771	0.9809	0.9838	0.9793	0.8840
seed = 2	0.9940	0.9771	0.9806	0.9839	0.9789	0.8837
seed = 3	0.9944	0.9771	0.9808	0.9838	0.9788	0.8839
seed = 4	0.9945	0.9771	0.9807	0.9839	0.9790	0.8840
ave	0.99432	0.97708	0.98078	0.98386	0.97904	0.88390
(std)	(0.00019)	(0.00004)	(0.00012)	(0.00005)	(0.00019)	(0.00011)

The results, summarized in Table 6, demonstrate that the uniqueness score for every model changes with the sample order. The standard deviation across the runs confirms this variability. These results reveal that the lack of permutation invariance is not just a theoretical flaw but a practical issue that can lead to inconsistent and unreliable model evaluations.