CRYSTAL GENERATIVE MODELING WITH EXPLICIT AUTOREGRESSIVE CONDITIONAL LIKELIHOODS AND NONTRIVIAL SPACE GROUP STABILIZERS

Rees Chang[1], **Alex Guerra**[2], **Nick Richardson**[2], **Ni Zhan**[2], **Sulin Liu**[3], **Angela Pak**[1], **Ryan Marr**[4], **Alex M. Ganose**[5], **Ryan P. Adams**[2*], **Elif Ertekin**[1*]
[1] University of Illinois at Urbana-Champaign, [2] Princeton University,
[3] Massachussets Institute of Technology, [4] University of Southern California,
[5] Imperial College London

ABSTRACT

Inverse crystalline materials design is a grand challenge in materials science. Most crystals have atoms at high-symmetry subspaces of 3D Euclidean space (i.e., positions with nontrivial stabilizer groups); yet, most existing crystal generative models cannot place atoms in these positions with nonzero probability. In this paper, we propose Wyckoff- and Asymmetric Unit-based Generative model (WyckoffAUGen), which sequentially builds crystals with explicit autoregressive-like conditional likelihoods and hard space group constraints. While prior methods parametrize distributions over unit cells with periodic translation symmetry, our model learns distributions over asymmetric units, which tile $\mathbb{R}^3$ upon applying the space group actions. This choice equips WyckoffAUGen with space group invariant model densities and reduces representations and generation trajectories to that of only symmetrically inequivalent atoms. To model continuous distributions over atom positions on facets of asymmetric units, WyckoffAUGen introduces a differentiable bijection from the simplex to any 2D polygon. Since experimental crystal synthesis can be hindered by unknown competing compounds in the same composition space, we enable masked in-filling from composition spaces.

1 INTRODUCTION

Crystals comprise critical technologies like batteries (Nitta et al., 2015), topological materials (Tang et al., 2019), electronic devices (Woods-Robinson et al., 2020), photovoltaics (Green et al., 2014), and more. Materials scientists have catalogued $\mathcal{O}(10^5)$ crystals experimentally (Bergerhoff et al., 1983) and $\mathcal{O}(10^6)$ *in silico* with density functional theory (DFT) simulation (Curtarolo et al., 2012; Jain et al., 2013; Saal et al., 2013). In contrast, the number of stable crystalline materials with five elements or less is estimated to exceed 10^{13}, and even higher-order compositions are common in real materials (Davies et al., 2016). Generative models offer a promising path to rapidly explore the vast space of crystals (Xie et al., 2022; Jiao et al., 2023; Miller et al., 2024; Cao et al., 2024).

Unlike molecules, crystals span the periodic table and exhibit discrete spatial symmetries according to one of 230 *space groups* (Aroyo et al., 2016). Specifically, crystals have invariances to discrete translations, rotations, reflections, and sequences thereof that transform atoms into themselves or into identical atoms. The list of space group actions which map a point into itself is called a *stabilizer group*. When a set of points in $\mathbb{R}^3$ have conjugate stabilizer groups, the set is called a *Wyckoff position*. Importantly, as shown in section A.1, Wyckoff positions can have zero volume, comprising points, lines, or planes. These zero volume sets are referred to as *special Wyckoff positions*. We give a more formal treatment of space groups and Wyckoff positions in section A.3.

Despite the fact that most existing crystal generative models ignore space groups and Wyckoff positions, they are critical for modeling real materials. Firstly, space groups and Wyckoff positions correlate strongly with materials properties; Neumann's principle states that all crystal properties

*Denotes equal advising

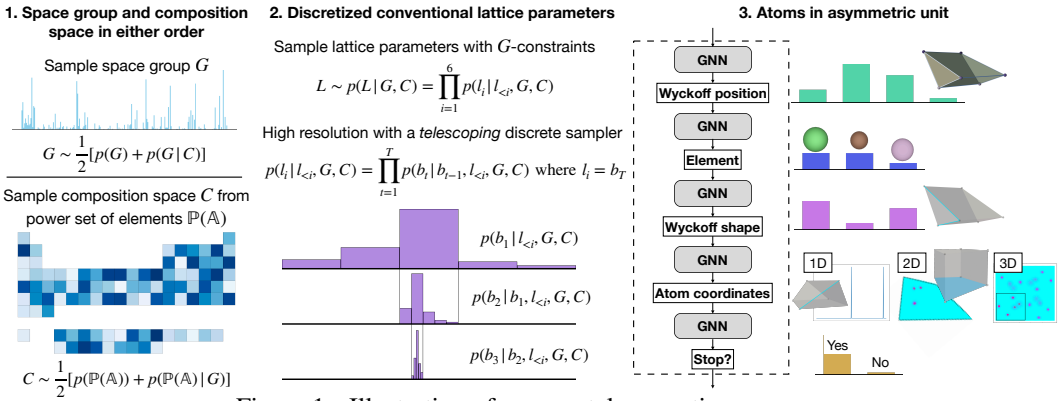

Figure 1: Illustration of our crystal generation process.

share the same invariances as the crystal itself (Neumann, 1885). Thus even slightly perturbing atoms out of special Wyckoff positions will reduce the crystal's space group symmetry and can subsequently cause significant (even discontinuous) changes to its macroscopic properties (Choi et al., 2009; Caretta et al., 2023; Shwartz et al., 2006; Cano & Bradlyn, 2020). Secondly, we show empirically in section A.1 that known materials usually occupy high symmetry space groups with atoms in zero-volume Wyckoff positions. Yet, most existing crystal generative models learn continuous distributions over all three spatial dimensions of atom positions, assigning zero probability measure to placing atoms in special Wyckoff positions. One might argue that crystals from these models can simply be relaxed into high symmetry positions with DFT or machine learned force fields. However, besides the apparent difficulty of generating atoms sufficiently close to high symmetry positions for relaxation (Zeni et al., 2025; Gruver et al., 2024), such a framework assigns different model probabilities to crystals that relax into the same structure, obfuscating training and evaluation.

Our contributions. In this paper, we introduce the development of WyckoffAUGen, a generative model that builds crystals sequentially, respecting hard constraints from space groups, Wyckoff positions, and composition spaces at every step. Our model (1) uses the symmetry-minimized representation of the *asymmetric unit* (ASU) and (2) trains with explicit, SE(3) and space group invariant, autoregressive conditional likelihoods. The model's sequential nature allows for masked in-filling, e.g., for hard-constrained sampling by space group for targeted search or by composition space for finding crystal phases hampering synthesis (McDermott et al., 2023; Doherty et al., 2021). We also introduce a method for learning explicit probability densities on arbitrary 2D polygons by leveraging generalized barycentric coordinates (Floater, 2003). For discussion on prior works, see A.2.

2 METHODOLOGY

Using `sympy` (Meurer et al., 2017) and `PyXtal` (Fredericks et al., 2021), we removed redundancy induced by space group symmetry from every Wyckoff position by intersecting each one with its exact asymmetric unit (see A.3 for details). Each zero-dimensional Wyckoff position was reduced to a single point, each one-dimensional position to a set of line segments, each two-dimensional position to a set of convex polygonal ASU facets, and each three-dimensional position to the asymmetric unit interior. The intersections for space group 192 are shown as an example in A.1.

We decomposed each crystal into $M = (G, \mathbb{C}, L, X, A, W) \in (\mathbb{G}, \mathbb{P}(\mathbb{A}), \mathbb{R}^{3\times3}, \mathbb{R}^{n\times3}, \mathbb{C}^n, \mathbb{W})$, where $G \in \mathbb{G}$ is the space group, $\mathbb{C} \in \mathbb{P}(\mathbb{A})$ is the composition space from the power set of elements $\mathbb{A}$ (see A.3), $L \in \mathbb{R}^{3\times3}$ is the conventional lattice basis, $A \in \mathbb{C}^n$ are elements, $X \in \mathbb{R}^{n\times3}$ are fractional atom coordinates, $W \in \mathbb{W}^n$ are Wyckoff positions, and n is the number of atoms in the asymmetric unit. Several variables in M impose hard constraints on each other. To handle these dependencies, the model sequentially samples each constrained variable after its constraining variable(s). We factorized the generation process as follows:

$$p(M) = \frac{1}{2}\left[p(\mathbb{C}|G)p(G) + p(G|\mathbb{C})p(\mathbb{C})\right] \times p(L|G,\mathbb{C}) \times \prod_{i=1}^{n}\Big[p(w_i|w_{<i},a_{<i},x_{<i},L,G,\mathbb{C}) \times p(a_i|w_{\leq i},a_{<i},x_{<i},L,G,\mathbb{C}) \times p(x_i|w_{\leq i},a_{\leq i},x_{<i},L,G,\mathbb{C})\Big] \quad (1)$$

2.1 SPACE GROUP AND COMPOSITION SPACE SAMPLING

We trained our model to sample space groups and composition spaces in either order to enable in-filling from either one at inference time. To model $p(G|\mathbb{C})$, we represented $\mathbb{C}$ as a one-hot vector of element occupancies and passed it through a multilayer perceptron (MLP) to produce space group logits. For modeling $p(G)$, we avoided the computational burden of marginalizing over all possible composition spaces by learning it separately as a simple length-230 vector of unnormalized logits. We factorized the discrete conditional distribution over composition spaces as $p(\mathbb{C}|G) = p(\text{stop}|c_{|\mathbb{C}|}) \prod_{i=1}^{|\mathbb{C}|} p(a_i|c_{<i}, G)$ where $c_i \in \mathbb{C}$, $a_i \in \mathbb{C} \setminus c_{<i}$, and $c_{|\mathbb{C}|} = \mathbb{C}$. To aid generalization across the 230 space groups, we formed one-hot features of G using the lattice centering type, crystal family, point group symbol, chirality, presence of inversion symmetry, and extra dimensions to differentiate collisions from screw and glide symmetries, yielding 62 features total. To predict $p(a_i|c_{<i}, G)$, we encoded $c_{<i}$ with a DeepSets model (Zaheer et al., 2017) on top of element embeddings $e_{\mathbb{A}}$ introduced by Xie & Grossman (2018). Similarly to $p(G)$, we separately learned $p(\mathbb{C})$ instead of marginalizing over the space groups, sharing parameters with $p(\mathbb{C}|G)$ by conditioning on a vector of zeroes.

2.2 TELESCOPING DISCRETE LATTICE SAMPLING

We parameterized univariate conditionals to autoregressively sample the 3 lattice lengths (a, b, c) and 3 angles between them (α, β, γ) for conventional unit cells, applying physical constraints to each conditional. Denoting $l = (a, b, c, \alpha, \beta, \gamma)$, the model learned

$$p(l|G, \mathbb{C}) = \prod_{i=1}^{6} p(l_i|l_{<i}, G, \mathbb{C}), \tag{2}$$

where $p(l_i|l_{<i}, G, \mathbb{C})$ has support over positive values with finite range determined by the data. Under the space group constraints, crystal lattices can be binned into 6 crystal families, each putting unique constraints on the lattice parameters. Space groups 1 to 2 impose no constraints; 3 to 15 require $\alpha = \gamma = 90°$; 16 to 74 require $\alpha = \beta = \gamma = 90°$; 75 to 142 require $\alpha = \beta = \gamma = 90°$ and $a = b$; 143 to 194 require $a = b$, $\alpha = \beta = 90°$, and $\gamma = 120°$; and 195 to 230 require $a = b = c$ and $\alpha = \beta = \gamma = 90°$. Furthermore, the crystal lattice must have non-zero volume. To impose these constraints, the model only learns univariate conditionals for the lattice parameters unconstrained by the crystal families, leaving constrained terms in the product of Equation 2 equal to 1. We enforce positive volume by dynamically setting the support of $p(\gamma|a, b, c, \alpha, \beta, G, \mathbb{C})$ to satisfy

$$\frac{\text{Volume}}{abc} = \sqrt{1 + 2\cos\alpha\cos\beta\cos\gamma - \cos^2\alpha - \cos^2\beta - \cos^2\gamma} > 0.$$

Besides physical constraints, lattice generation requires the flexibility to learn highly peaked distributions since small perturbations to a crystal lattice can significantly alter materials properties. In $BaTiO_3$ for example, 0.03Å strain was found to increase the ferroelectric transition temperature by 500°C and the remnant polarization by 250% (Choi et al., 2004). In contrast, the range of conventional lattice lengths in the MP20 dataset is over 100Å. To address this challenge, we chose to discretize the lattice parameters to a resolution of 0.01Å. Naively, this resolution requires a softmax over $N_l = \mathcal{O}(10^4)$ classes per lattice parameter to achieve a 100Å range. We overcame this poor scaling by *telescoping* the categorical distribution. See Figure 1 for a visual explanation. At a high level, the range of lattice parameters was first binned very coarsely, and a class b_1 was sampled from $p(b_1|l_{<i}, G, \mathbb{C})$. Then, the selected class b_1 was further coarsely binned and one of these higher resolution bins was selected from $p(b_2|b_1, l_{<i}, G, \mathbb{C})$. This process was repeated T times to achieve higher levels of resolution as

$$p(l_i|l_{<i}, G, \mathbb{C}) = \prod_{t=1}^{T} p(b_t|b_{t-1}, l_{<i}, G, \mathbb{C})$$

where $l_i = b_T$ and $b_0 = \emptyset$. Choosing $p(b_t|b_{t-1}, l_{<i}, G, \mathbb{C})$ to be a categorical distribution over a small number of classes $n_l \ll N_l$ achieves $\frac{1}{(n_l)^T} = \frac{1}{N_l}$ resolution with $\mathcal{O}(n_l T) \ll \mathcal{O}(N_l)$ memory. We used $T = 2$ and $n_l = 100$ in our experiments. We represented conditioning on l_i and b_t using random Fourier features (Tancik et al., 2020) and produced the logits for $p(b_t|b_{t-1}, l_{<i}, G, \mathbb{C})$ with an MLP.

2.3 ATOM SAMPLING

Our model samples atoms in the asymmetric unit one at a time, conditioning on a graph neural network (GNN)-predicted embedding h_i of G, $\mathbb{C}$, L, and all previously generated atoms. To sample atom coordinates, the model first samples a convex set of points where the Wyckoff position intersects the ASU (see A.3.3). We refer to each convex set as a *Wyckoff shape*. Finally, the model samples continuous coordinates from a probability density with support on the Wyckoff shape, which may be zero-, one-, two-, or three-dimensional. After sampling atom i, the model may terminate generation by sampling from a Bernoulli distribution $p(\text{stop}|w_{\leq i}, a_{\leq i}, x_{\leq i}, L, G, \mathbb{C})$. Further details on the model architecture and training are in A.4.

Wyckoff positions and elements. To learn a categorical distribution $p(w_i|w_{<i}, a_{<i}, x_{<i}, L, G, \mathbb{C})$ over Wyckoff positions for atom i, we used the GNN to produce an embedding h_i of the crystal so far, concatenated embeddings $e_{\mathbf{w}}$ of Wyckoff positions allowed to be sampled from the space group, and predicted logits for each of these Wyckoff positions. For this purpose, unique 231-dimensional feature vectors were created for the 1731 Wyckoff positions across all space groups using Wyckoff multiplicities, Wyckoff dimensionalities, site symmetry symbols, the space group features from Section 2.1, and average Fourier features of the coordinates at vertices and centers of masses of the Wyckoff shapes. We set logits of zero-dimensional Wyckoff positions which are already occupied to negative infinity to prevent sampling overlapping atoms at those locations. Similarly, we parameterized a categorical distribution $p(a_i|w_{\leq i}, a_{<i}, x_{<i}, L, G, \mathbb{C})$ over elements by taking element embeddings $e_a \in e_{\mathbb{A}}$ and predicting logits as $\text{MLP}(e_a||h_i||e_{w_i})$, where $||$ denotes concatenation. We masked logits of elements outside the composition space, i.e., $\mathbb{A} \setminus \mathbb{C}$, to negative infinity.

Wyckoff shapes. We found that the intersections of one- and two-dimensional Wyckoff positions with the asymmetric unit may consist of multiple line segments or convex polygonal facets. We placed categorical distributions over these shapes $s_i \in \mathbb{S}$ as $p(s_i|w_{\leq i}, a_{\leq i}, x_{<i}, L, G, \mathbb{C})$. Each shape's embedding e_{s_i} was predicted with a DeepSets model on random Fourier features of the shape vertices' conventional cell fractional coordinates. The logit for each shape $s_{i'}$ was produced as $\text{MLP}(e_{s_{i'}}||h_i||e_{w_i}||e_{a_i})$.

Coordinates in 1D Wyckoff shapes. We placed a mixture of Beta distributions on each one-dimensional "Wyckoff line segment" as $p(x_i|s_i, w_{\leq i}, a_{\leq i}, x_{<i}, L, G, \mathbb{C})$, where the mixture model parameters were predicted as $\text{MLP}(h_i||e_{w_i}||e_{a_i}||e_{s_i})$.

Coordinates in 2D Wyckoff shapes. To model a probability density on a convex polygonal facet P_k of an asymmetric unit, we parameterized Dirichlet mixture distributions $p_{\text{Dirichlet}}(\cdot)$ on the triangle $T_3 \in \mathbb{R}^2$. We leveraged generalized barycentric coordinates (see A.3) to bijectively map T_3 to P_k with vertices $V^{P_k} \in \mathbb{R}^{k\times 2}$. To create the map, we first pretended that the triangle T_3 is a k-gon T_k with vertices $V^{T_k} \in \mathbb{R}^{k\times 2}$ by appending $(k-3) \geq 0$ non-vertex points to the boundary ∂T_3. Then, given a point $y \in T_3$ and denoting $\phi^{T_k}(y) \in \mathbb{R}^k$ as the generalized barycentric coordinates of y with respect to T_k, we constructed the map $f : T_3 \to P_k$ as $f(y) = \phi^{T_k}(y)V^{P_k}$. The probability density of a point $x \in P_k$ is then given by the change of variables formula (Rezende & Mohamed, 2015). We pre-computed T_k for every 2D Wyckoff shape by minimizing the distortion (Eq. 3) of f with respect to V^{T_k} evaluated at quadrature points from `Basix` (Scroggs et al., 2022). Dirichlet mixture model parameters were predicted in the same manner as the 1D Beta mixture parameters.

Coordinates in 3D Wyckoff positions. We modeled probability densities in 3D Wyckoff positions by learning a probability density $p_\Gamma(x)$ in the conventional unit cell Γ and wrapping it around the asymmetric unit Π. Specifically, for $x \in \Pi$ and space group G, we parameterized $p_\Pi(x) = \sum_{g\in G, gx\in\Gamma} p_\Gamma(gx)$, where $p_\Gamma(\cdot)$ was a mixture of von Mises distributions whose parameters were predicted in the same manner as the 1D Beta mixture parameters.

2.4 OBJECTIVE FUNCTION

Since there are potentially many construction orderings σ leading to the same crystal, we reduced the number of orderings by enforcing lexicographic partial orderings on both the data and model. Atoms were ordered lexicographically by Wyckoff letter, then atomic number. Composition spaces were ordered lexicographically by atomic number. The ordering of remaining variables were sampled uniformly randomly at each training step. Similarly to Uria et al. (2014), we maximized a lower

Table 1: Results on the MP20 dataset.

	Sampling time ↓	Validity (%) ↑		U.N. rate (%) ↑	Distribution distance ↓				CMD ↓	Diversity ↑	
	(sec / batch)	Structure	Composition	$\mathcal{D}_{\text{train}}^{\text{MP20}}$	W_ρ	$W_{N_{el}}$	JSD_G	$\text{JSD}_{d_{\text{Wyckoff}}}$	Structure	Structure	Composition
CDVAE	906	**99.99**	85.66	**98.2**	0.6590	1.423	0.6957	0.4590	0.4821	0.6539	13.70
DiffCSP	154	99.92	82.21	85.6	**0.1454**	0.4000	0.4638	0.2328	0.1766	0.9588	**15.69**
DiffCSP++	484	99.92	85.94	84.7	0.1658	0.5002	**0.1608**[*]	**0.0449**[*]	**0.1079**	0.9329	15.23
SymmCD	139	88.24	**86.76**	87.7	0.1640	0.3213	**0.1669**[*]	**0.0344**	0.3233	0.9111	15.62
FlowMM (reported)	-	96.85	83.19	-	-	-	-	-	-	-	-
WyckoffAUGen	**2**	82.35	82.20	34.3	0.3367	**0.0379**	0.2752	0.1071	0.2225	**0.9614**	15.53

[*] Uses fixed templates from the training data.

bound on the likelihood as

$$\log p(M) = \log \mathbb{E}_{\sigma \sim \Pr(\sigma)} \Big[p(M_\sigma) \Big] \geq \mathbb{E}_{\sigma \sim \Pr(\sigma)} \Big[\log p(M_\sigma) \Big].$$

3 Results and Discussion

We evaluated WyckoffAUGen on MP20 (Xie et al., 2022), a benchmark dataset of real materials. Evaluations were conducted on a budget of 10,000 generated crystals. We computed structural and compositional validity percentages using heuristics about interatomic distances and charge, respectively. We note that only ∼90% of real crystals in the MP20 dataset pass the compositional validity checker based on `SMACT` Davies et al. (2016) and thus should be assessed with caution. Of the 10,000 generated crystals, we randomly sampled 1,000 which were determined to be both structurally and compositionally valid. Of these 1,000 crystals, we determined how many were unique and novel (*U.N*) with respect to the training dataset using `pymatgen`'s `StructureMatcher` (Ong et al., 2013) with `stol=0.3`, `angle_tol=5`, and `ltol=0.2`. U.N. crystals were used to compute (1) distribution distances between ground truth test and generated materials properties, including Wasserstein distances for atomic density ρ and number of unique elements N_{el} as well as Jensen-Shannon divergences for space group G and occupied Wyckoff dimensionalities d_{Wyckoff}; (2) Central Moment Discrepancy (CMD) (Zellinger et al., 2017) up to 50 moments between ground truth test and generated crystal CrystalNN structural fingerprints (Zimmermann & Jain, 2020); and (3) structural and compositional diversity as measured by average pairwise L_2-distances between CrystalNN and Magpie (Ward et al., 2016) fingerprints, respectively. We also measured average sampling times per batch of 500 crystals on a single NVIDIA A40 GPU.

Evaluation metrics are in Table 1. Unsurprisingly, WyckoffAUGen strongly outperformed CDVAE and DiffCSP on space group and Wyckoff dimensionality metrics. WyckoffAUGen also generates crystals ∼70x faster than DiffCSP and ∼220x faster than DiffCSP++. WyckoffAUGen performs competitively on property distance and diversity metrics but underperforms on sampling structurally valid crystals. The U.N. rate of WyckoffAUGen is lower than other models, but a more useful metric also computes *stability* (Zeni et al., 2025) which we intend to pursue. We show non-cherry picked random and U.N. crystals generated by WyckoffAUGen in Figures 4 and 5, respectively.

The mixture models used to generate atom coordinates in WyckoffAUGen enable rapid sampling and scalable likelihood-based training with explicit probabilities. These may be useful, e.g., to estimate thermodynamic ensemble properties with Boltzmann generators (Noé et al., 2019; Volokhova et al., 2024), approximate Bayesian posteriors, or train GFlowNets (Bengio et al., 2021). However, the mixture models only directly condition on previously generated atoms, in contrast to score-based diffusion which also conditions on atom locations currently being denoised. We hypothesize that this may explain WyckoffAUGen's subpar performance on the structural validity metric.

4 Conclusion and Future Work

We built an autoregressive model with explicit space group invariant likelihoods enabling hard constrained infilling from space groups, composition spaces, or otherwise incomplete crystals. Limitations of the current approach include underperforming in generating structurally valid crystals; producing 1- and 2-dimensional densities over atom coordinates discontinuously over periodic boundaries; and ignoring crystal classes like molecular, magnetic, and 2D crystals. We plan to explore alternative approaches to sampling atom positions, conduct DFT-based evaluations, and pursue property-guided generation. Leveraging emerging paradigms from large language models like test-time compute scaling (Snell et al., 2024) may also be an interesting avenue for future work.

ACKNOWLEDGMENTS

This research used the Delta advanced computing and data resource (award OAC 2005572) and the Illinois Campus Cluster, operated by the Illinois Campus Cluster Program in conjunction with the National Center for Supercomputing Applications. The research was supported by the National Science Foundation under Grant Nos. DGE 21-46756 and 2118201.

REFERENCES

Ryan P. Adams and Peter Orbanz. Representing and Learning Functions Invariant Under Crystallographic Groups, 2023. URL http://arxiv.org/abs/2306.05261.

M.I. Aroyo, H. Burzlaff, G. Chapuis, W. Fischer, H.D. Flack, A.M. Glazer, H. Grimmer, B. Gruber, Th. Hahn, H. Klapper, E. Koch, P. Konstantinov, V. Kopsky, D.B. Litvin, A. Looijenga-Vos, K. Momma, U. Mueller, U. Shmueli, B. Souvignier, J.C.H. Spence, P.M. de Wolff, H. Wondratschek, and H. Zimmerman. *International Tables for Crystallography: Space-group symmetry*, volume A. International Union of Crystallography, Chester, England, 2 edition, 2016. ISBN 978-0-470-97423-0. doi: 10.1107/97809553602060000114. URL https://it.iucr.org/Ac/.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=Arn2E4IRjEB.

G. Bergerhoff, R. Hundt, R. Sievers, and I. D. Brown. The Inorganic Crystal Structure Data Base. *Journal of Chemical Information and Computer Sciences*, 23(2):66–69, 1983. ISSN 00952338. doi: 10.1021/ci00038a003.

Jennifer Cano and Barry Bradlyn. Band Representations and Topological Quantum Chemistry. *Annual Review of Condensed Matter Physics*, 12:225–246, 2020. doi: 10.1146/annurev-conmatphys-041720-124134. URL http://arxiv.org/abs/2006.04890http://dx.doi.org/10.1146/annurev-conmatphys-041720-124134.

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

Lucas Caretta, Yu-Tsun Shao, Jia Yu, Antonio B Mei, Bastien F Grosso, Cheng Dai, Piush Behera, Daehun Lee, Margaret McCarter, Eric Parsonnet, Harikrishnan K. P, Fei Xue, Xiangwei Guo, Edward S Barnard, Steffen Ganschow, Zijian Hong, Archana Raja, Lane W Martin, Long-Qing Chen, Manfred Fiebig, Keji Lai, Nicola A Spaldin, David A Muller, Darrell G Schlom, and Ramamoorthy Ramesh. Non-volatile electric-field control of inversion symmetry. *Nature Materials*, 22(2):207–215, 2023. ISSN 1476-4660. doi: 10.1038/s41563-022-01412-0. URL https://doi.org/10.1038/s41563-022-01412-0.

K J Choi, M Biegalski, Y L Li, A Sharan, J Schubert, R Uecker, P Reiche, Y B Chen, X Q Pan, V Gopalan, L.-Q. Chen, D G Schlom, and C B Eom. Enhancement of Ferroelectricity in Strained BaTiO3 Thin Films. *Science*, 306(5698):1005–1009, 2004. doi: 10.1126/science.1103218. URL https://www.science.org/doi/abs/10.1126/science.1103218.

T Choi, S Lee, Y J Choi, V Kiryukhin, and S.-W. Cheong. Switchable Ferroelectric Diode and Photovoltaic Effect in BiFeO₃. *Science*, 324(5923):63–66, 2009. doi: 10.1126/science.1168636. URL https://www.science.org/doi/abs/10.1126/science.1168636.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret

Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. URL http://jmlr.org/papers/v24/22-1144.html.

Stefano Curtarolo, Wahyu Setyawan, Shidong Wang, Junkai Xue, Kesong Yang, Richard H Taylor, Lance J Nelson, Gus L W Hart, Stefano Sanvito, Marco Buongiorno-Nardelli, Natalio Mingo, and Ohad Levy. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Computational Materials Science*, 58:227–235, 2012. doi: 10.1016/j.commatsci.2012.02.002.

Daniel W Davies, Keith T Butler, Adam J Jackson, Andrew Morris, Jarvist M Frost, Jonathan M Skelton, and Aron Walsh. Computational Screening of All Stoichiometric Inorganic Materials. *Chem*, 1(4):617–627, 2016. ISSN 2451-9294. doi: https://doi.org/10.1016/j.chempr.2016.09.010. URL https://www.sciencedirect.com/science/article/pii/S2451929416301553.

Max Welling Diederik P. Kingma. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Tiarnan A.S. Doherty, Satyawan Nagane, Dominik J. Kubicki, Young Kwang Jung, Duncan N. Johnstone, Affan N. Iqbal, Dengyang Guo, Kyle Frohna, Mohsen Danaie, Elizabeth M. Tennyson, Stuart Macpherson, Anna Abfalterer, Miguel Anaya, Yu Hsien Chiang, Phillip Crout, Francesco Simone Ruggeri, Sean M. Collins, Clare P. Grey, Aron Walsh, Paul A. Midgley, and Samuel D. Stranks. Stabilized tilted-octahedra halide perovskites inhibit local formation of performance-limiting phases. *Science*, 374(6575):1598–1605, 2021. ISSN 10959203. doi: 10.1126/SCIENCE.ABL4890. URL https://www.science.org/doi/10.1126/science.abl4890.

Alexandre Duval, Victor Schmidt, Alex Hernandez Garcia, Santiago Miret, Fragkiskos D Malliaros, Yoshua Bengio, and David Rolnick. FAENet: Frame Averaging Equivariant GNN for Materials Modeling. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 9013–9033, 2023.

Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Michael S. Floater. Mean value coordinates. *Computer Aided Geometric Design*, 20(1):19–27, 2003. ISSN 0167-8396. doi: 10.1016/S0167-8396(03)00002-5.

Scott Fredericks, Kevin Parrish, Dean Sayre, and Qiang Zhu. PyXtal: A Python library for crystal structure generation and symmetry analysis. *Computer Physics Communications*, 261:107810, 2021. ISSN 0010-4655. doi: https://doi.org/10.1016/j.cpc.2020.107810. URL https://www.sciencedirect.com/science/article/pii/S0010465520304057.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In Z Ghahramani, M Welling, C Cortes, N Lawrence, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper{_}files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Martin A. Green, Anita Ho-Baillie, and Henry J. Snaith. The emergence of perovskite solar cells. *Nature Photonics 2014 8:7*, 8(7):506–514, 2014. ISSN 1749-4893. doi: 10.1038/nphoton.2014.134. URL https://www.nature.com/articles/nphoton.2014.134.

R. W. Grosse-Kunstleve, N. K. Sauter, and P. D. Adams. Numerically stable algorithms for the computation of reduced unit cells. *Acta Cryst.*, 60(1):1–6, 2004. ISSN 0108-7673. doi: 10.1107/S010876730302186X. URL https://journals.iucr.org/paper?sh5006https://journals.iucr.org/a/issues/2004/01/00/sh5006/.

Ralf W. Grosse-Kunstleve, Buddy Wong, Marat Mustyakimov, and Paul D. Adams. Exact direct-space asymmetric units for the 230 crystallographic space groups. *Acta Crystallogr A*, 67(3): 269–275, 2011. ISSN 0108-7673. doi: 10.1107/S0108767311007008. URL //journals.iucr.org/paper?pz5088.

Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C. Lawrence Zitnick, and Zachary Ward Ulissi. Fine-tuned language models generate stable inorganic materials as text. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vN9fpfqoP1.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In H Larochelle, M Ranzato, R Hadsell, M F Balcan, and H Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper{_}files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

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

Anubhav Jain, Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A Persson. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater*, 1:11002, 2013. doi: 10.1063/1.4812323. URL https://doi.org/10.1063/1.4812323.

Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal Structure Prediction by Joint Equivariant Diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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. URL https://openreview.net/forum?id=jkvZ7v4OmP.

Nikita Kazeev, Ruiming Zhu, Ignat Romanov, Andrey E Ustyuzhanin, Shuya Yamazaki, Wei Nong, and Kedar Hippalgaonkar. Wyckofftransformer: Generation of symmetric crystals. In *AI for Accelerated Materials Design - NeurIPS 2024*, 2024. URL https://openreview.net/forum?id=Jcy1bPOqrY.

Sungwon Kim, Juhwan Noh, Geun Ho Gu, Alán Aspuru-Guzik, and Yousung Jung. Generative Adversarial Networks for Crystal Structure Prediction. *ACS Central Science*, 6(8):1412–1420, 2020. URL http://arxiv.org/abs/2004.01396.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. URL http://arxiv.org/abs/1412.6980.

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. URL https://openreview.net/forum?id=xnssGv9rpW.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Matthew J McDermott, Brennan C McBride, Corlyn E Regier, Gia Thinh Tran, Yu Chen, Adam A Corrao, Max C Gallant, Gabrielle E Kamm, Christopher J Bartel, Karena W Chapman, Peter G Khalifah, Gerbrand Ceder, James R Neilson, and Kristin A Persson. Assessing Thermodynamic Selectivity of Solid-State Reactions for the Predictive Synthesis of Inorganic Materials. *ACS*

Central Science, 9(10):1957–1975, 2023. ISSN 2374-7943. doi: 10.1021/acscentsci.3c01051. URL https://doi.org/10.1021/acscentsci.3c01051.

Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, 2017. ISSN 2376-5992. doi: 10.7717/peerj-cs.103. URL https://doi.org/10.7717/peerj-cs.103.

Benjamin Kurt Miller, Ricky T.Q. Chen, Anuroop Sriram, and Brandon M. Wood. FlowMM: Generating Materials with Riemannian Flow Matching. *Proceedings of Machine Learning Research*, 235:35664–35686, 2024. ISSN 26403498. URL https://arxiv.org/abs/2406.04713v1.

Mistal, Alex Hernández-García, Alexandra Volokhova, Alexandre AGM Duval, Yoshua Bengio, Divya Sharma, Pierre Luc Carrier, Michał Koziarski, and Victor Schmidt. Crystal-GFN: sampling materials with desirable properties and constraints. In *AI for Accelerated Materials Design - NeurIPS 2023 Workshop*, 2023. URL https://openreview.net/forum?id=l167FjdPOv.

F.E. Neumann. *Vorlesungen über die Theorie der Elasticität der festen Körper und des Lichtäthers, gehalten an der Universität Königsberg*. Leipzig, G. G., 1885.

Naoki Nitta, Feixiang Wu, Jung Tae Lee, and Gleb Yushin. Li-ion battery materials: present and future. *Materials Today*, 18(5):252–264, 2015. ISSN 1369-7021. doi: 10.1016/J.MATTOD.2014.10.040.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457), 2019. ISSN 10959203. doi: 10.1126/SCIENCE.AAW1147. URL https://www.science.org/doi/10.1126/science.aaw1147.

Juhwan Noh, Jaehoon Kim, Helge S Stein, Benjamin Sanchez-Lengeling, John M Gregoire, Alan Aspuru-Guzik, and Yousung Jung. Inverse Design of Solid-State Materials via a Continuous Representation. *Matter*, 1:1370–1384, 2019. doi: 10.1016/j.matt.2019.08.017. URL https://doi.org/10.1016/j.matt.2019.08.017.

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. ISSN 09270256. doi: 10.1016/j.commatsci.2012.10.028.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Zekun Ren, Juhwan Noh, Siyu Tian, Felipe Oviedo, Guangzong Xing, Qiaohao Liang, Armin Aberle, Yi Liu, Qianxiao Li, Senthilnath Jayavelu, Kedar Hippalgaonkar, Yousung Jung, and Tonio Buonassisi. Inverse design of crystals using generalized invertible crystallographic representation. *Matter*, 5(1), 2022. URL https://arxiv.org/abs/2005.07609.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1530–1538, Lille, France, 2015. PMLR. URL https://proceedings.mlr.press/v37/rezende15.html.

James E Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and C Wolverton. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM*, 65:1501–1509, 2013. doi: 10.1007/s11837-013-0755-4.

Lisa Schneckenreiter, Richard Freinschlag, Florian Sestak, Johannes Brandstetter, Günter Klambauer, and Andreas Mayr. GNN-VPA: A variance-preserving aggregation strategy for graph neural networks. In *5th Workshop on practical ML for limited/low resource settings*, 2024. URL https://openreview.net/forum?id=cmrRmu2afD.

Matthew W. Scroggs, Igor A. Baratta, Chris N. Richardson, and Garth N. Wells. Basix: a runtime finite element basis evaluation library. *Journal of Open Source Software*, 7(73):3982, 2022. doi: 10.21105/joss.03982. URL https://doi.org/10.21105/joss.03982.

Sharon Shwartz, Raoul Weil, Mordechai Segev, Eugene Lakin, Emil Zolotoyabko, Vinod M Menon, Stephen R Forrest, and Uri El-Hanany. Light-induced symmetry breaking and related giant enhancement of nonlinear properties in CdZnTe:V crystals. *Opt. Express*, 14(20):9385–9390, 2006. doi: 10.1364/OE.14.009385. URL https://opg.optica.org/oe/abstract.cfm?URI=oe-14-20-9385.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. URL https://arxiv.org/abs/2408.03314.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Yang Song and Stefano Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. In *Advances in Neural Information Processing Systems*, 2019.

Anuroop Sriram, Benjamin Kurt Miller, Ricky T. Q. Chen, and Brandon M Wood. FlowLLM: Flow matching for material generation with large language models as base distributions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=0bFXbEMz8e.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. In H Larochelle, M Ranzato, R Hadsell, M F Balcan, and H Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7537–7547. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper{_}files/paper/2020/file/55053683268957697aa39fba6f231c68-Paper.pdf.

Feng Tang, Hoi Chun Po, Ashvin Vishwanath, and Xiangang Wan. Comprehensive search for topological materials using symmetry indicators. *Nature*, 566(7745):486–489, 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-0937-5. URL https://www.nature.com/articles/s41586-019-0937-5.

Atsushi Togo, Kohei Shinohara, and Isao Tanaka. Spglib: a software library for crystal symmetry search. *Science and Technology of Advanced Materials: Methods*, 4:1, 2024. doi: 10.1080/27660400.2024.2384822. URL https://www.tandfonline.com/doi/abs/10.1080/27660400.2024.2384822.

Benigno Uria, Iain Murray, and Hugo Larochelle. A deep and tractable density estimator. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 467–475, Bejing, China, 2014. PMLR. URL https://proceedings.mlr.press/v32/uria14.html.

Alexandra Volokhova, Michał Koziarski, Alex Hernández-García, Cheng Hao Liu, Santiago Miret, Pablo Lemos, Luca Thiede, Zichao Yan, Alán Aspuru-Guzik, and Yoshua Bengio. Towards equilibrium molecular conformation generation with GFlowNets. *Digital Discovery*, 3(5):1038–1047, 2024. ISSN 2635098X. doi: 10.1039/D4DD00023D. URL https://pubs.rsc.org/en/content/articlehtml/2024/dd/d4dd00023dhttps://pubs.rsc.org/en/content/articlelanding/2024/dd/d4dd00023d.

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):16028, 2016. ISSN 2057-3960. doi: 10.1038/npjcompumats.2016.28. URL https://doi.org/10.1038/npjcompumats.2016.28.

Rachel Woods-Robinson, Yanbing Han, Hanyu Zhang, Tursun Ablekim, Imran Khan, Kristin A. Persson, and Andriy Zakutayev. Wide Band Gap Chalcogenide Semiconductors. *Chemical Reviews*, 120(9):4007–4055, 2020. ISSN 15206890. doi: 10.1021/ACS.CHEMREV.9B00600. URL https://pubs.acs.org/doi/full/10.1021/acs.chemrev.9b00600.

Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120:145301, 2018. doi: 10.1103/PhysRevLett.120.145301. URL https://link.aps.org/doi/10.1103/PhysRevLett.120.145301.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal Diffusion Variational Autoencoder for Periodic Material Generation. In *International Conference on Learning Representations*, 2022.

Keqiang Yan, Xiner Li, Hongyi Ling, Kenna Ashen, Carl Edwards, Raymundo Arroyave, Marinka Zitnik, Heng Ji, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. Invariant tokenization of crystalline materials for language model enabled generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=18FGRNd0wZ.

Sherry Yang, Simon Batzner, Ruiqi Gao, Muratahan Aykol, Alexander L Gaunt, Brendan McMorrow, Danilo J Rezende, Dale Schuurmans, Igor Mordatch, and Ekin D Cubuk. Generative hierarchical materials search. *arXiv preprint arXiv:2409.06762*, 2024a.

Sherry Yang, KwangHwan Cho, Amil Merchant, Pieter Abbeel, Dale Schuurmans, Igor Mordatch, and Ekin Dogus Cubuk. Scalable diffusion for materials generation. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=wm4WlHoXpC.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep Sets. In I Guyon, U Von Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper{_}files/paper/2017/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf.

Werner Zellinger, Edwin Lughofer, Susanne Saminger-Platz, Thomas Grubinger, and Thomas Natschläger. Central Moment Discrepancy (CMD) for Domain-Invariant Representation Learning. In *International Conference on Learning Representations*, 2017. URL http://www.scch.at.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Aliaksandra Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, 2025. URL https://arxiv.org/abs/2312.03687v2.

Ruiming Zhu, Wei Nong, Shuya Yamazaki, and Kedar Hippalgaonkar. WyCryst: Wyckoff inorganic crystal generator framework. *Matter*, 7(10):3469–3488, 2024. ISSN 2590-2393. doi: 10.1016/j.matt.2024.05.042. URL https://doi.org/10.1016/j.matt.2024.05.042.

Nils E.R. 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. ISSN 20462069. doi: 10.1039/c9ra07755c. URL https://pubs.rsc.org/en/content/articlehtml/2020/ra/c9ra07755chttps://pubs.rsc.org/en/content/articlelanding/2020/ra/c9ra07755c.

A Appendix

A.1 Motivating Wyckoff positions and asymmetric units

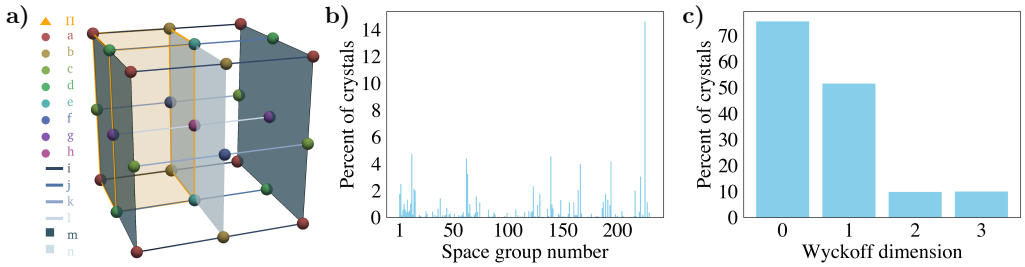

Figure 2: (a) The asymmetric unit (Π) and special Wyckoff positions labeled by letter in the conventional unit cell of space group 10. (b-c) Histograms of occupied space groups and Wyckoff dimensionalities by crystals in the MP20 (Xie et al., 2022; Jain et al., 2013; Bergerhoff et al., 1983) training dataset. Space groups and Wyckoff positions were determined by the `SpaceGroupAnalyzer` module in `pymatgen` (Ong et al., 2013; Togo et al., 2024) using tolerances of 0.1 Å and 5°. These tolerances help account for the moderate convergence criteria of the Materials Project DFT relaxations.

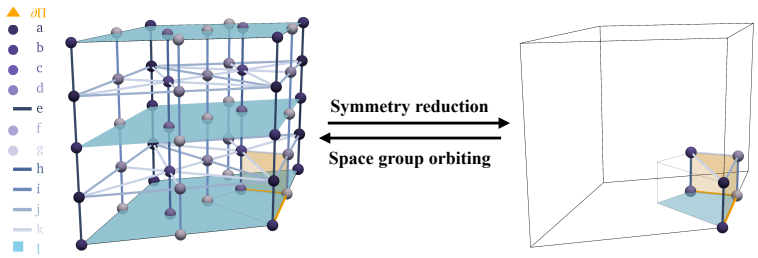

Figure 3: Special Wyckoff positions and the asymmetric unit in the conventional unit cell of hexagonal space group 192. Closed asymmetric unit boundary edges and facets (∂Π) are in orange.

A.2 Related work

Early crystal generative models represented crystals as voxelized images (Noh et al., 2019; Hoffmann et al., 2019) or padded tensors of 3D coordinates (Kim et al., 2020; Ren et al., 2022) to train variational autoencoders (VAE) (Diederik P. Kingma, 2014) or generative adversarial networks (Goodfellow et al., 2014). Recent works have enforced the SE(3) and periodic translational invariance of crystals by leveraging graph neural networks. One popular approach is to use diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020) on crystal lattices, atom types, and atom positions (Xie et al., 2022; Jiao et al., 2023; Zeni et al., 2025). These models have also been extended with the flow matching framework (Lipman et al., 2023), accelerating sampling (Miller et al., 2024). Other works have attempted to learn the SE(3) and periodic invariances through data augmentations (Yang et al., 2024b; Gruver et al., 2024) or data canonicalization (Yan et al., 2024) with image diffusion (Yang et al., 2024b) or large language models (LLMs) (Gruver et al., 2024; Yan et al., 2024). Concurrent works have used LLMs to generate noisy crystals which are then refined with graph-based diffusion or flow matching (Sriram et al., 2024; Yang et al., 2024a).

Two of the aforementioned works attempted to learn space group-conditioned generation without hard constraints. The graph diffusion model MatterGen (Zeni et al., 2025) was fine-tuned on 14 space groups and used ground truth numbers of atoms per unit cell per space group to initialize generation. However, they could only generate target space groups with 20% accuracy as assessed by `pymatgen`'s `SpaceGroupAnalyzer` using unreported tolerance values (Ong et al., 2013; Togo et al., 2024). Similarly, CrystalLLM (Gruver et al., 2024) only managed 24% accuracy despite a generous `SpaceGroupAnalyzer` tolerance of 0.2Å.

More relevant to our work, a few other models have considered hard space group constraints during generation. WyCryst (Zhu et al., 2024) trained a VAE to generate atom types and Wyckoff position occupations, but rely on post-hoc DFT calculations to relax atom positions from uniformly random locations in the Wyckoff positions. Crystal-GFN (Mistal et al., 2023) considered space group constraints for the task of distribution matching under the GFlowNet framework (Bengio et al., 2021) but did not address how to sample atom coordinates with space group constraints. DiffCSP++ (Jiao et al., 2024) trained a graph-based diffusion model with masked diffusion of the unit cell lattice, continuous element diffusion with a post-hoc `argmax`, and projected diffusion of atom positions on the Wyckoff subspaces. They achieved space group invariance by averaging the denoising term over all atoms in a unit cell belonging to the same Wyckoff position. However, DiffCSP++ does not readily emit likelihoods due to the post-hoc `argmax`; and they do not learn to sample space groups, numbers of atoms per unit cell, or Wyckoff position occupations, instead relying on templates from the training data. CrystalFormer (Cao et al., 2024) trained a transformer-based autoregressive model, canonicalizing crystals as a sequence of atoms ordered lexicographically by Wyckoff letter and then fractional coordinates. The model was trained to learn these orderings and sample atom coordinates in special Wyckoff positions by conditionally masking an amortized mixture of 3D von Mises distributions. However, their von Mises distributions are not space group invariant, thus erroneously assigning different likelihoods to symmetrically equivalent atoms. Concurrently to our work, SymmCD (Levy et al., 2025) and WyckoffTransformer (Kazeev et al., 2024) also consider space group constrained generation. SymmCD is a diffusion model which leverages asymmetric units to reduce memory footprints, but uses discrete diffusion of Wyckoff positions and elements and then post-hoc projections of atomic coordinates to satisfy Wyckoff position constraints; thus SymmCD cannot readily yield explicit likelihoods. WyckoffTransformer predicts atom types and Wyckoff positions but relies on interatomic potentials (which do not preserve space group symmetry) or DiffCSP++ to determine atom coordinates and thus also cannot readily yield explicit likelihoods. Unlike these existing works, our model learns to generate crystals from scratch; produces explicit space group invariant, AR-like conditional likelihoods; parametrizes distributions over Wyckoff positions in asymmetric units instead of unit cells; and learns distributions over composition spaces.

A.3 Preliminaries

A.3.1 Space groups

Formally, a space group $G \in \mathbb{G}$ is a group of isometries that tiles $\mathbb{R}^3$ with a convex polytope Π called the *asymmetric unit* (ASU) (Adams & Orbanz, 2023; Grosse-Kunstleve et al., 2011). In particular, G is generated by an infinite subgroup of discrete lattice translations $T = \{n_1\mathbf{l}_1, n_2\mathbf{l}_2, n_3\mathbf{l}_3 | n_i \in \mathbb{Z}, \mathbf{l}_i \in \mathbb{R}^3\}$ as well as a collection of other symmetry operations $g(\cdot) = \{R(\cdot) + v | R \in \mathrm{O}(3), v \in \mathbb{R}^3\} \in G$, where R is a point group operation (rotation, reflection, or identity) and v is a translation.

A.3.2 Wyckoff positions

Given a space group and a point $x \in \mathbb{R}^3$, the *stabilizer* group $G_x \coloneqq \{g | gx = x\} \subset G$ is the finite subgroup of G that leaves x invariant. A Wyckoff position is then defined as the set of points with conjugate stabilizer groups, i.e., $\{x' | \exists\ g \in G : G_{x'} = gG_xg^{-1}\}$. Conceptually, if g is a point group operation, this means that all points in a Wyckoff position are invariant to the same space group operations up to a change of basis. By convention, when x is described with respect to the lattice basis $\{\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3\}$, the size of the *orbit* of x in the unit cell, $|\{gx | g \in G, gx \in [0, 1)^3\}|$, is called the Wyckoff *multiplicity*. Wyckoff positions whose stabilizer groups are non-trivial, i.e., include more than the identity operation, are referred to as *special Wyckoff positions* as opposed to the *general Wyckoff position* defined by the identity stabilizer group. Wyckoff positions are labeled by multiplicity and *Wyckoff letter*, where the lexicographic ordering gives the Wyckoff positions in order of increasing multiplicity.

A.3.3 Representing crystals with the asymmetric unit

Previous crystal generative models represent the infinite translational periodicity of a crystal with a parallelipiped Γ called the *unit cell*. The unit cell reduces infinite crystals by removing redundancy induced by T, the group of discrete lattice translations. In this way, crystals are represented by the tuple $M = (A, X, L)$, where $A = (a'_1, ..., a'_N) \in \mathbb{A}^N$ are the atom types, $\mathbb{A}$ is the set of all

chemical elements, and N is the number of atoms in the unit cell; $X = \{(x'_1, ..., x'_N)|x'_i \in \Gamma\}$ are the Cartesian atom coordinates; and $L = (L_1, L_2, L_3) \in \mathbb{R}^{3\times3}$ are the unit cell basis vectors. Given M, the infinite periodic structure can be reconstructed by applying the actions of T as $\{(a'_i, x'_i + n_1L_1 + n_2L_2 + n_3L_3)_{i=1}^N | n_j \in \mathbb{Z}\}$. Alternatively, the atom coordinates can be given in the lattice basis instead of the Cartesian basis. In this case, the infinite crystal is reconstructed as $\{\big(a'_i, x'_i + n_1e_1 + n_2e_2 + n_3e_3\big)_{i=1}^N | n_j \in \mathbb{Z}\}$. For the rest of this paper, we assume atom coordinates are always in the lattice basis. While the choice of unit cell is not unique, prior crystal generative models (Xie et al., 2022; Jiao et al., 2023; Miller et al., 2024) either canonicalize it with a minimum-volume *primitive cell* determined by the Niggli algorithm (Grosse-Kunstleve et al., 2004) or (Gruver et al., 2024; Cao et al., 2024; Jiao et al., 2024) a *conventional cell* which contains all the symmetries of the space group (Aroyo et al., 2016).

The pitfall of the unit cell representation is that, for 229 of the 230 space groups, it contains atoms which are symmetrically equivalent. Thus unit cell-based generative models which independently introduce even minute errors into atom positions will break any space group symmetry that is not a lattice translation.

In our work, we represented crystals with a convex polytope $\Pi \in \mathbb{R}^3$, the ASU, which maximally reduces infinite crystals by removing all redundancies induced by the space group $G \supseteq T$. Under this formulation, we consider atoms in the ASU with fractional coordinates $X = \{(x_1, ..., x_n)|x_i \in \Pi\} \in \mathbb{R}^{n\times3}$, atom types $A = (a_1, ..., a_n) \in \mathbb{A}^n$, and Wyckoff positions $W = \{(w_1, ..., w_n)|w_i = G_{x_i}\} \in \mathbb{W}^n$ where $n \leq N$. The infinite periodic structure of a crystal can be reconstructed by applying the actions of G to Π, i.e.,

$$\{(a_i, g_{ij}x_i) \mid x_i \in \Pi,\ g_{ij} \in G/w_i,\ i \in (1, ..., n)\}.$$

By only considering these symmetrically inequivalent atoms, we reduce our model's memory footprint and minimize the dimensionality of the generative modeling task. Restricting our model probability distributions to the ASU also makes them automatically space group invariant. We canonicalize the non-unique choice of ASU using those listed in the International Tables for Crystallography (Aroyo et al., 2016) with additional conditions on faces, edges, and vertices from Grosse-Kunstleve et al. (2011) to ensure that the ASUs are *exact*, i.e., that Π tiles $\mathbb{R}^3$ without overlaps at the boundaries $\partial\Pi$.

A.3.4 Composition space

A composition space contains all possible stoichiometries that can be formed using a subset of elements from the periodic table. Constraining a generative model to a composition space with fewer elements than the periodic table is practically relevant for informing materials synthesis experiments wherein only certain elements are allowed to enter the reaction chamber (McDermott et al., 2023; Doherty et al., 2021). We represent a *composition space* as an unordered set of elements $\mathbb{C} \in \mathbb{P}(\mathbb{A})$ from the power set over all elements $\mathbb{A}$ in the periodic table. A given crystal resides in any composition space that is a superset of the crystal's elements, leading to a combinatorial explosion of composition spaces that can be identified with a crystal. In practice, we restrict the maximum size of our composition spaces to 7 elements, the maximum found in the MP20 dataset.

A.3.5 Generalized barycentric coordinates

Given an arbitrary polygon $P \subset \mathbb{R}^2$ with vertices $V \in \mathbb{R}^{k\times2}$ ordered counterclockwise and $k \geq 3$, *generalized barycentric coordinates* (Floater, 2003) are defined as the function $\phi : P \to \mathbb{R}^k$, which, for all $x \in P$, satisfies

$$\phi(x)_i \geq 0, \quad \mathbf{1}^T\phi(x) = 1, \quad \phi(x)V = x$$

where $\mathbf{1} \in \mathbb{R}^k$ is the ones vector. When $k = 3$, the generalized barycentric coordinates are uniquely determined as the usual barycentric coordinates of a triangle. For $k > 3$, the choice is no longer unique. We chose the "mean value coordinates" (Floater, 2003) defined as the smooth functions,

$$\phi(x)_i = \frac{w_i(x)}{\sum_{j=1}^{k} w_j(x)}, \quad w_i = \frac{\tan(\alpha_{i-1}/2) + \tan(\alpha_i/2)}{||v_i - x||}$$

where $\alpha_i = \angle V_i x V_{i+1} \in (0, \pi)$ and $V_{k+1} \equiv V_0$.

When mapping triangles to polygons with $f : T_3 \to P_k$, we calculated the distortion of the map as

$$\text{Distortion} = \frac{1}{N_{\text{quad}}} \sum_{i=1}^{N_{\text{quad}}} \left(\left| \frac{\partial f^{-1}(x_i; V^{T_k})}{\partial x_i} \right| - 1 \right)^2 \tag{3}$$

where x_i is a quadrature point.

A.4 Architecture and Training

Our code was written with PyTorch Paszke et al. (2017) and PyTorch Geometric (Fey & Lenssen, 2019). The model was trained with the AdamW optimizer (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) on a single NVIDIA A100 GPU. Our GNN was a modified version of FAENet (Duval et al., 2023), replacing sum pooling with variance-preserving aggregation (Schneckenreiter et al., 2024) and removing frame averaging since we trivially achieve $\mathrm{SE}(3)$ and space group invariance by canonicalizing crystals with the ASU representation. We constructed fully connected atom graphs $\mathcal{G}$ wherein each atom in the primitive unit cell was connected to every other atom in the primitive unit cell by their minimum-length distance and relative positions under periodic boundary conditions. If ties existed, all corresponding edges were included. For improved memory, only node embeddings of atoms in the asymmetric unit were computed. Rescaling by Wyckoff multiplicities was employed at pooling operations to maintain consistency with unit cell representations. Seed nodes with special learnable embeddings were placed at the origin to represent crystals without any atoms. The architecture is summarized as follows:

$$u_i^0 \leftarrow \mathrm{MLP}\big(e_{w_i} || e_{a_i} || \mathrm{FourierEmbedding}(x_i)\big) \tag{4}$$

$$e_{ij} \leftarrow \frac{1}{m_i} \mathrm{MLP}\big(\hat{\mathrm{r}}_{\mathrm{ij}} || \mathrm{RBF}(\mathrm{d}_{\mathrm{ij}})\big) \tag{5}$$

$$f_{ij}^k \leftarrow \mathrm{MLP}\big(e_{ij} || u_i^k || u_j^k\big) \tag{6}$$

$$u_i^{k+1} \leftarrow u_i^k + a \cdot \mathrm{MLP} \circ \mathrm{GraphNorm}\left(\frac{1}{\sqrt{|\mathcal{N}_{\mathrm{i}}|}} \sum_{\mathrm{j} \in \mathcal{N}_{\mathrm{i}}} \mathrm{u}_{\mathrm{j}}^{\mathrm{k}} \odot \mathrm{f}_{\mathrm{ij}}^{\mathrm{k}} \right) \tag{7}$$

$$u_i^{k_{\max}} \leftarrow \mathrm{MLP}(u_i^0 || ... || u_i^{k_{\max}}) \tag{8}$$

$$g \leftarrow \frac{\sum_i^n m_i \cdot \alpha(u_i^{k_{\max}}) \cdot u_i^{k_{\max}}}{\sqrt{\sum_i^n \left(m_i \cdot \alpha(u_i^{k_{\max}}) \right)^2}} \tag{9}$$

$$\tag{10}$$

$$h \leftarrow \mathrm{MLP}\big(g || e_G || e_{\mathbb{C}} || e_\rho || e_l\big) \tag{11}$$

where u_i^k is the node feature of the ith atom in the asymmetric unit after k rounds of message passing; $\hat{r}_{ij} = \frac{\vec{r}_{ij}}{||\vec{r}_{ij}||}$ is the normalized Cartesian relative position between atoms i and j; d_{ij} is the pairwise Cartesian distance; $k_{\max}$ is the number of message passing layers; m_i is the Wyckoff multiplicity of the ith atom; a is a learnable scalar initialized to zero; $\mathrm{FourierEmbedding}(x_i)$ are random Fourier features of atom i's fractional coordinates under the conventional unit cell lattice basis; n is the number of atoms in the asymmetric unit; $\alpha(\cdot)$ are learnable attention weights; e_{a_i} is the embedding of atom i's element; e_{w_i} is the embedding of atom i's Wyckoff position; e_G, $e_{\mathbb{C}}$, e_ρ, and e_l are embeddings of the crystal's space group, composition space, atomic density, and lattice parameters, respectively; and h is the final crystal embedding.

A.4.1 Stabilization techniques

- We prevented the variance of probability densities over atom coordinates from going to zero by constraining maximum mixture component distribution parameters with a $-\mathrm{softplus}(\cdot)$ operation.
- To better align the distribution of partially complete crystals seen during training and inference, we employed noisy teacher forcing during training. Specifically, we maximized the probability of ground truth atoms and lattice parameters conditioned on noisy ground truth

atoms and lattice parameters. We applied isotropic Gaussian noise with a standard deviation of 0.2Å to atom positions, restricting the noise to the subspace of the atom's Wyckoff position and applying the periodic boundary conditions of the ASU whenever noise moved atoms outside the ASU. Lattice lengths and angles were augmented with noise from uniform distributions with 0.01Å and 1° ranges, respectively, and then discretized following 2.2.

- To ensure various component probability distributions converged at similar rates during training, gradients were re-balanced with straight-through estimators as

$$\log p_i \leftarrow c \log p_i - \text{stop\_grad}(c \log p_i) + \text{stop\_grad}(\log p_i).$$

  We set $c = 0.3$ for sampling lattice parameters, Wyckoff positions, elements, and termination and $c = 1.0$ otherwise.
- To prevent mixture models with K modes from collapsing to $k \ll K$ modes that dominate the gradient signal, we did *maximum a posteriori* estimation by placing a Dirichlet prior over mixture weights with $\alpha_{\text{Dirichlet}} = (1.0001)^K$.
- We adopted auxiliary z-loss regularization from Chowdhery et al. (2023) with $\lambda = 10^{-4}$ to prevent the log normalizers of Wyckoff, element, and termination logits from getting too large.

A.4.2 HYPERPARAMETERS

Hyperparameter	Value
GNN learning rate	2×10^{-4}
Space group learning rate	1×10^{-3}
Composition space learning rate	2×10^{-3}
Lattice learning rate	1×10^{-3}
Atoms learning rate	5×10^{-4}
Weight decay	10^{-5}
AdamW ϵ	10^{-5}
Gradient clipping value	1.0
Batch size	512
Epochs	3000
Hidden dimension	256
Relative position filters	480
Relative distance embeddings	300
Message passing layers	7
Beta distribution max α, β	5000
Dirichlet distribution max α	2500
von Mises distribution max κ	500
Beta mixture components	10
Dirichlet mixture components	30
Von Mises mixture components	50
Lattice length Fourier scale	20.0
Lattice angle Fourier scale	1.0
Lattice parameter bin edges Fourier scale	10.0
Number of parameters	8,968,929

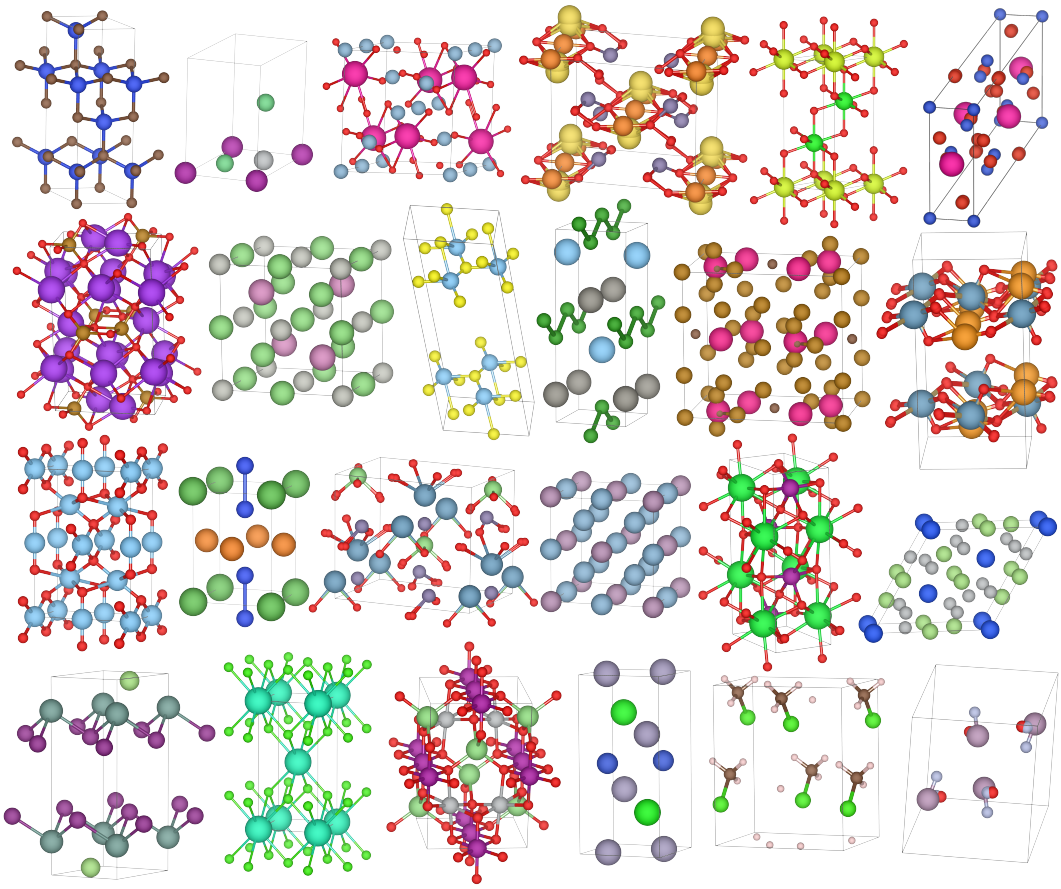

Figure 4: Non-cherry picked random crystals generated by WyckoffAUGen.

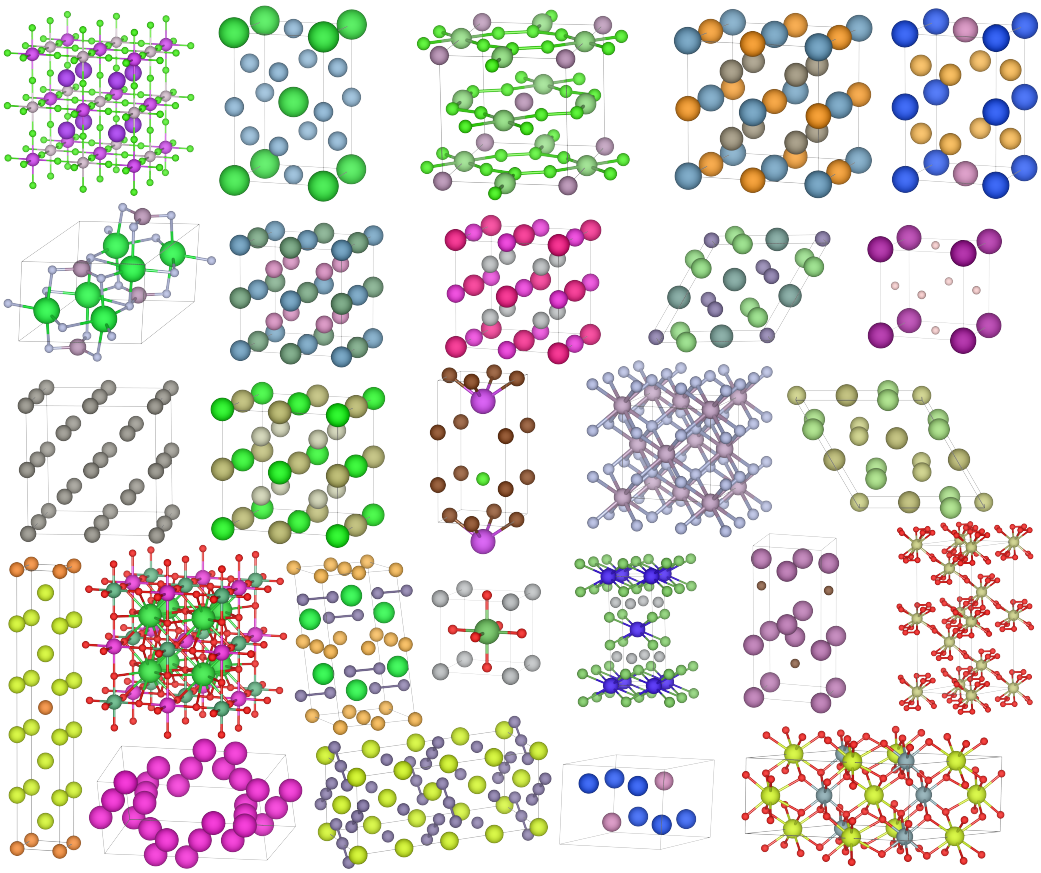

Figure 5: Non-cherry picked unique and novel (with respect to the MP20 training data) crystals generated by WyckoffAUGen. The prominence of Heusler crystals exposes the bias of the MP20 training dataset.