WyckoffTransformer: Generation of Symmetric Crystals

Nikita Kazeev
Institute for Functional Intelligent Materials
National University of Singapore
Singapore 117544
kna@nus.edu.sg

Ruiming Zhu
School of Materials Science and Engineering
Nanyang Technological University
Singapore 639798
raymond_zhurm@outlook.com

Ignat Romanov
Tsinghua University
Haidian District, 100084, Beijing, China
linty24@mails.tsinghua.edu.cn

Andrey Ustyuzhanin
Constructor University Bremen gGmbH
Campus Ring 1, Bremen, 28759, Germany
andrey.u@gmail.com

Shuya Yamazaki
School of Materials Science and Engineering
Nanyang Technological University
Singapore 639798
shuya001@e.ntu.edu.sg

Wei Nong
School of Materials Science and Engineering
Nanyang Technological University
Singapore 639798
NONG0003@e.ntu.edu.sg

Kedar Hippalgaonkar
School of Materials Science and Engineering
Nanyang Technological University
Singapore 639798
kedar@ntu.edu.sg

Abstract

We propose WyckoffTransformer, a generative model for inorganic materials that takes advantage of the high order symmetry present in most known crystals. Wyckoff positions, a mathematical object from space group theory, is used as the basis for an elegant, compressed, and discrete structure representation. To model the distribution we develop a permutation–invariant autoregressive model based on Transformer. Our experiments demonstrate that Wyckoff Transformer has better performance compared to the baseline in generating novel stable structures conditioned on the space group symmetry, while also having competitive metric values when compared to a model not conditioned on space group symmetry.

1 Introduction

Space of all possible combinations of atoms forming periodic structures is intractably large. It is not possible to screen it fully, even with a fast machine learning algorithm. Practical materials, however, occupy only a small part of it [1]. Firstly, they must correspond to an energy minimum. Secondly, occupying an energy minimum is not sufficient to establish if the material is synthesizable or indeed

[1] Conjecture: the set of all possible stable materials is countable, while the set of all possible ways to place atoms in 3D space has cordiality of the continuum. (Dis)proving it goes beyond this paper.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

experimentally stable. Having a generative model that outputs a priori stable materials is a step towards speeding up automated material design by orders of magnitude.

2 Space groups and Wyckoff positions

Our work relies on a crucial insight: most ($\approx 98\%$) known crystals have symmetry beyond the simplest, $P1$, which only requires lattice translation (see also figure 4). Those symmetries are not merely a mathematical observation; optical, electrical, magnetic, structural and other properties are determined by symmetry [15, 21].

Wyckoff positions (WPs) are regions of space that are invariant under some transformation. Mathematically, a WP is any point in a set of points whose site symmetry groups are all conjugate subgroups one of another [14]. Every point inside a unit cell belongs to a WP. A note on terminology: an "atom position" is just a single point in 3D space, while a "Wyckoff position" is a set of 3D points. Some WPs contain a finite number of 3D points, while other WPs correspond to 1D lines, 2D planes, or an open continuous 3D set. See figure 1 for an illustration. WPs for a given space group are commonly enumerated by Latin letters in the order of multiplicity and denoted by a combination of the multiplicity value and the letter, e. g. `2a`.

A crystal can be represented as a space group and a set of WPs and chemical elements occupying them, reducing the number of parameters by an order of magnitude without information loss. For example, see figure 2.

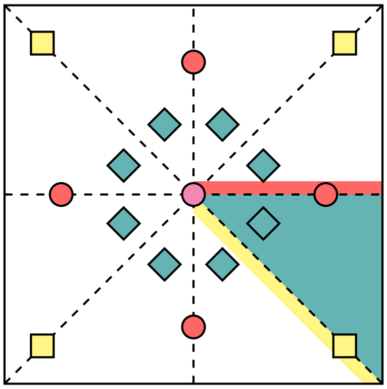

Figure 1: Wyckoff positions illustration of a toy 2D crystal [8]. It contains 4 mirror lines, and one rotation center. There are four Wyckoff positions, illustrated by shading. Magenta is the Wyckoff position that is invariant under all the transformations, it only contains a single point; red and yellow lie on the mirror lines, and teal is only invariant under identity transformation and occupies the rest of the space. Markers of the corresponding colors show one of the possible locations of an atom belonging to the corresponding Wyckoff position.

```
Group: I4/mmm (139)
Lattice: a = b = 8.9013, c = 5.1991, α = 90.0, β = 90.0, γ = 90.0
Wyckoff sites:
Nd @ [ 0.0000  0.0000  0.0000], WP [2a] Site [4/m2/m2/m]
Al @ [ 0.2788  0.5000  0.0000], WP [8j] Site [mm2.]
Al @ [ 0.6511  0.0000  0.0000], WP [8i] Site [mm2.]
Cu @ [ 0.2500  0.2500  0.2500], WP [8f] Site [..2/m]
```

Figure 2: Wyckoff representation of Nd(Al_2Cu)$_4$ (mp-974729), variable parameters in **bold**. If represented as a point cloud, the structure has 13[atoms]$\times$3[coordinates]$+$6[lattice] $=$ 42 parameters; if represented using WPs, it has just 4 continuous parameters (WPs `8i` and `8j` each have a free parameter, and the tetragonal lattice has two), and 5 discrete parameters (space group number, and WPs for each atom).

3 WyckoffTransformer

3.1 Tokenization

A complete crystal representation consists of two parts: discrete, containing space group, WPs and chemical elements; and continuous, containing fractional coordinates and lattice parameters. In this paper, we are only concerned with the discrete part. Given a Wyckoff representation, coordinates can be determined with random initialization followed by a relaxation with a machine learning potential [24], or DiffCSP++ [13].

We represent each structure as a set of tokens, as shown in figure 3. The first token contains the space group, the others chemical elements and WPs. We encode a WP as a tuple containing site symmetry and so–called *enumeration*. Several WPs can correspond to the isomorphic site symmetries, for example both yellow and red in figure 1 correspond to reflection, but with different axes. To differentiate those WPs we enumerate them separately within each space group and site symmetry according to conventional WP order [3]. For example, in space group 225 present in figure 3 WP `4a` becomes `(m-3m, 0)`, `4b` becomes `(m-3m, 1)`, and `8c` becomes `(-43m, 0)`. The purpose of this encoding is to take advantage of the fact that, unlike Wyckoff letters, site symmetry definition is universal across different space groups.

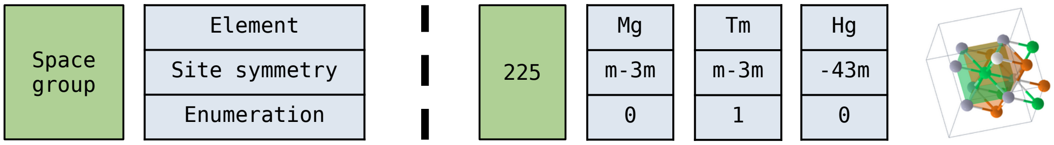

Figure 3: An example of structure tokenization, $TmMgHg_2$ mp-865981

3.2 Model architecture

Token embeddings are constructed by concatenating the embeddings for every part of the token (element, site symmetry, enumeration). To each structure we add a `STOP` token. We use those sequences as an input for an encoder–only Transformer [19]. We do not use positional encoding. To represent states where some parts of token are known and others are not, we replace those values with `MASK`. We take the output of Transformer on the token containing `MASK` value(s), concatenate it with a one–hot vector encoding presence of each token in the input sequence, and use the result as the input for three fully–connected neural networks, one for each predicted part of token.

3.3 Training

We predict the next token in cascade: first the chemical element, then, conditioned on it, site symmetry and, finally, enumeration. We use cross–entropy loss. Validation dataset is used for learning rate scheduling and early stopping.

Wyckoff representation is permutation–invariant, so is Transformer. Auto–regressive generation, however, is not permutation–invariant. Therefore we shuffle the order of every Wyckoff representation at every training epoch and use multi–class loss when training the element prediction. See Appendix D for analysis.

For some crystals *enumerations* part of Wyckoff representation is not uniquely defined, as it depends on the arbitrary choice of the coset representative of the space group affine normalizer. For example the structure figure 3 has two equivalent representations: [(Mg, `(m-3m,` **`0`**`)`), (Tm, `(m-3m,` **`1`**`)`), (Hg, `(-43m, 0)`)] and [(Mg, `(m-3m,` **`1`**`)`), (Tm, `(m-3m,` **`0`**`)`), (Hg, `(-43m, 0)`)]. Fortunately, the number of variants is small; in MP-20 [20] dataset for 96% structures there are less than 10. We encourage invariance also using data augmentation, by picking a randomly selected equivalent representation at every training epoch.

3.4 Structure generation

Generation is conditioned on space group number which is sampled from the combination of training and validation datasets. Wyckoff representation is then autoregressively sampled using the WyckoffTransformer. Finally we generate a crystal structure using pyXtal [7] and relax it with CHGNet [6], see Appendix E. As an alternative option, we also used DiffCSP++[13].

4 Experimental Evaluation

The goal of the generative model is to produce novel stable structures from a distribution approximating the training data. We use a multi–step procedure for evaluation of the generated structures. Firstly, we discard non–unique generated structures. For all models the number is ≤ 3 out of 1000, so uniqueness is not discussed in detail. Secondly, we count and discard structures that are present in the training and validation dataset and thus not novel, as verified by `pymatgen.analysis.StructureMatcher` [17] with default parameters. Thirdly, we evaluate stability and statistical similarity between the generated structures and the test dataset. Stability is approximated by energy above convex hull $E_{\text{hull}} < 0.08$ eV, the same threshold as used during construction of MP-20 dataset; the criterion is computed by CHGNet [6] with Materials Project [11] as the source of the structures for the hull. For consistency, we do not use DFT energies for reference structures, but recompute energy with CHGNet.

We train and evaluate WyckoffTransformer on MP-20 dataset [20] (45229 structures), following the original 60/20/20 train/validate/test split. Since WyCryst [24] only supports a limited number of chemical elements per structure, to compare with it WyckoffTransformer was separately trained on a subset of MP-20 containing only binary and ternary structures (35575 in total). All structures are relaxed with CHGNet.

WyckoffTransformer is better than DiffCSP and WyCryst in reproducing property statistics; 39% of the structures generated by WyckoffTransformer are stable unique and novel (S.U.N) [23], which is less than for DiffCSP 57.4%. It is greater then the value reported for MatterGen-MP [23, fig. 2e] ($\approx 25\%$), but they cannot be compared directly as in that work DFT is used for energy computation. WyckoffTransformer S.U.N. fraction is similar to WyCryst, but the later has much lower novelty, so it is likely that the WyCryst–generated structures that do not verbatim appear in the training data are still similar to it. Structures generated by WyckoffTransformer by design belong to the target space group; only 20% do for MatterGen [23, sec. 2.4]. Detailed comparison is present in table 1, and plots are available in appendix C.

Model	**WyckoffTransformer**	DiffCSP	WyCryst	Test dataset
	Dataset: MP-20			
Novelty % ↑	**90**	**90**	-	100
Charge neutrality %	80	**81**	-	91
Num sites KS ↓	**0.058**	0.415	-	0
Num elements KS ↓	**0.054**	0.163	-	0
DoF KS ↓	**0.067**	0.407	-	0
$P1$ %	**3**	37	-	2
Space group χ^2 ↓	**0.222**	7.95	-	0
Elements χ^2 ↓	**0.040**	0.100	-	0
S.U.N. % ↑	39.2 pyXtal / 36.7 DiffCSP++	**57.4**	-	-
	Dataset: MP-20 binary & ternary			
Formal validity % ↑	**97**	-	72	-
Novelty % ↑	**89**	-	53	100
Charge neutrality %	**79**	-	76	90
Num sites KS ↓	**0.052**	-	0.287	0
Num elements KS ↓	0.030	-	**0.002**	0
DoF KS ↓	**0.056**	-	0.282	0
$P1$ %	**1**	-	5	1
Space group χ^2 ↓	**0.063**	-	0.82	0
Elements χ^2 ↓	**0.019**	-	0.140	0
S.U.N. % ↑	**38.7** pyXtal	-	36.7	-

Table 1: Evaluation of WyckoffTransformer, WyCryst and DiffCSP. All metrics below Novelty are computed only using novel structures. Detailed definitions are available in appendix B.

5 Related work

Crystal generation is a burgeoning field, with a plethora of models operating in the 3D point cloud space [12, 4, 22, 23]. Large Language Models that use a straightforward text representation of structures as CIF [10] have been recently proposed [1, 9, 16]. Our approach complements them naturally, by providing symmetry-based constraints and / or initial structure approximation.

Our work is inspired by [24], the first generative model to utilize Wyckoff positions. Our primary contribution is an autoregressive token-based model, as opposed to a VAE, allowing for a better inductive bias, and production of materials with a varying number of elements. A recent preprint [4] independently explores a similar approach.

Wyckoff Transformer improves on other machine learning works that use WPs in a few key ways. 1. Previous works encode WPs by enumerating them by space group defined letters [24, 8, 4, 2], while we use universally-defined site symmetries, easing transfer learning. 2. It explicitly combines all the information about a WP into a single token, while [4] relies on the sequential order to maintain the relationship between the chemical element and WP. 3. Instead of imposing an arbitrary order [4], we train our model to work for all permutations 4. We are also the first ML work to take into account the dependency of Wyckoff representation on the arbitrary choice of the coset representative of the space group affine normalizer.

6 Limitations and Conclusion

Our work shares the limitation of most generative models for materials [23, 24, 20, 12]: we learn the distribution from the training dataset, so there must be stable structures that are out-of-domain and will not be generated. E_{hull} as a proxy for stability is commonly used, but is imperfect, as it does not take into account entropy, and the hull determination relies on known structures. The high number of asymmetric space group $P1$ structures found by DiffCSP, that are supposedly stable, casts additional doubt on stability estimation with CHGNet E_{hull}.

Novelty evaluation is crucial. A model can generate structures that are same or similar to the ones in the training dataset, and are valid, but not very useful for material design. Counting complete duplicates is a step in the right direction, but does not measure substantial sample diversity.

WyckoffTransformer has a strong inductive bias motivated by the fact that most of known crystals are highly symmetric. It introduces several architecture innovations to create a deep learning model best suited for this representation. The distribution of structures generated by WyckoffTransformer is closer to the test dataset than that for DiffCSP and WyCryst in terms of the high–level property statistics, with a larger fraction of stable structures than WyCryst.

Acknowledgments and Disclosure of Funding

We thank Andrey Okhotin for the proposed metric of permutation invariance.

This research/project is supported by the Ministry of Education, Singapore, under its Research Centre of Excellence award to the Institute for Functional Intelligent Materials (I-FIM, project No. EDUNC-33-18-279-V12). This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-RP-2022-028). The computational work for this article was performed on resources at the National Supercomputing Centre of Singapore (NSCC) and NUS HPC. The research used computational resources provided by Constructor Tech.

References

[1] Nawaf Alampara, Santiago Miret, and Kevin Maik Jablonka. MatText: Do Language Models Need More than Text & Scale for Materials Modeling? *arXiv preprint arXiv:2406.17295*, 2024.

[2] Nada Ali Hakami and Hanan Ahmed Hosni Mahmoud. Deep learning classification of crystal structures utilizing wyckoff positions. *Crystals*, 12(10):1460, 2022.

[3] Mois Ilia Aroyo, Juan Manuel Perez-Mato, Cesar Capillas, Eli Kroumova, Svetoslav Ivantchev, Gotzon Madariaga, Asen Kirov, and Hans Wondratschek. Bilbao crystallographic server: I. databases and crystallographic computing programs. *Zeitschrift für Kristallographie-Crystalline Materials*, 221(1):15–27, 2006.

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

[5] Daniel W Davies, Keith T Butler, Adam J Jackson, Jonathan M Skelton, Kazuki Morita, and Aron Walsh. Smact: Semiconducting materials by analogy and chemical theory. *Journal of Open Source Software*, 4(38):1361, 2019.

[6] Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J Bartel, and Gerbrand Ceder. Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, 2023.

[7] 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.

[8] Rhys EA Goodall, Abhijith S Parackal, Felix A Faber, and Rickard Armiento. Wyckoff set regression for materials discovery. In *Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020), Vancouver, Canada.*, 2020.

[9] 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.

[10] Sydney Hall and Brian McMahon. *International tables for crystallography, volume G: Definition and exchange of crystallographic data*. John Wiley & Sons, 2005.

[11] 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.

[12] 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, 2024.

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

[14] Lev Kantorovich. *Quantum theory of the solid state: an introduction*, volume 136. Springer Science & Business Media, 2004.

[15] Cécile Malgrange, Christian Ricolleau, and Michel Schlenker. *Symmetry and physical properties of crystals*. Springer, 2014.

[16] Santiago Miret and NM Krishnan. Are llms ready for real-world materials discovery? *arXiv preprint arXiv:2402.05200*, 2024.

[17] 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.

[18] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[19] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[20] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *ICLR 2022, arXiv preprint arXiv:2110.06197*, 2021.

[21] Jiashi Yang et al. *An introduction to the theory of piezoelectricity*, volume 9. Springer, 2005.

[22] Mengjiao Yang, KwangHwan Cho, Amil Merchant, Pieter Abbeel, Dale Schuurmans, Igor Mordatch, and Ekin Dogus Cubuk. Scalable diffusion for materials generation, 2023.

[23] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha 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. MatterGen: a generative model for inorganic materials design, 2024.

[24] Ruiming Zhu, Wei Nong, Shuya Yamazaki, and Kedar Hippalgaonkar. WyCryst: Wyckoff inorganic crystal generator framework. *Matter*, 2024.

Appendix

A Space groups distribution

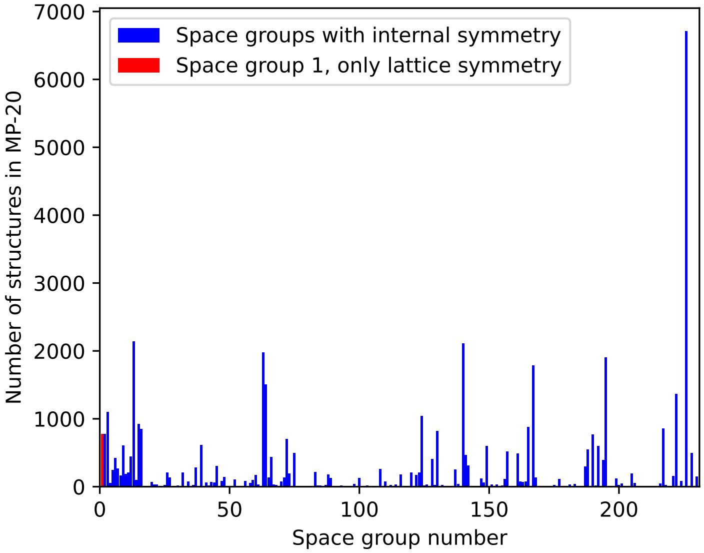

Figure 4: Distribution of space groups in MP-20 dataset [20]. 98% of structures belong to symmetry groups other that $P1$, with internal symmetries.

B Property comparison details

We do not use COV metrics [20], as they are almost saturated; they reach values $\geq 99.5\%$ for state-of-the-art methods [20, 12, 22] with the maximum being 100%.

For DiffCSP we use the generated structures provided by the authors. DiffCSP++ was retrained.

Formal validity is a fraction of formally consistent Wyckoff representations. A Wyckoff representations is invalid if it contains site symmetries not present in the space group, enumerations not valid for a given site symmetry – space group combination, or several atoms are forced into a single point in 3D space by being assigned to the same WP with no degrees of freedom.

Novelty is defined as the fraction of generated structures that do not appear in the training dataset, as verified by pymatgen `StructureMatcher` with default parameters.

Charge neutrality is computed using SMACT [5]. Only 91% of structures in MP-20 are charge neutral, so it is not a necessary criterion for structure validity and stability.

Num sites KS is the Kolmogorov–Smirnov statistic between the number of Wyckoff sites in the generated and test datasets. See also figures 6, 9.

Num elements KS is the Kolmogorov–Smirnov statistic between the number of unique chemical elements in the generated and test datasets. See also figures 5, 8.

DoF KS is the Kolmogorov–Smirnov statistic between the number degrees of freedom of Wyckoff representations between the generated and test datasets.

Space group χ^2 is the χ^2 statistic of difference of the frequencies of space groups between the generated and test datasets.

Elements χ^2 is the χ^2 statistic of difference of the frequencies of different elements between the generated and test datasets.

S.U.N. is the fraction of novel unique structures having $E_{\text{hull}} < 0.08$ eV computed with CHGNet. See also figures 7, 10.

C Property plots

C.1 WyckoffTransformer vs DiffCSP; MP-20

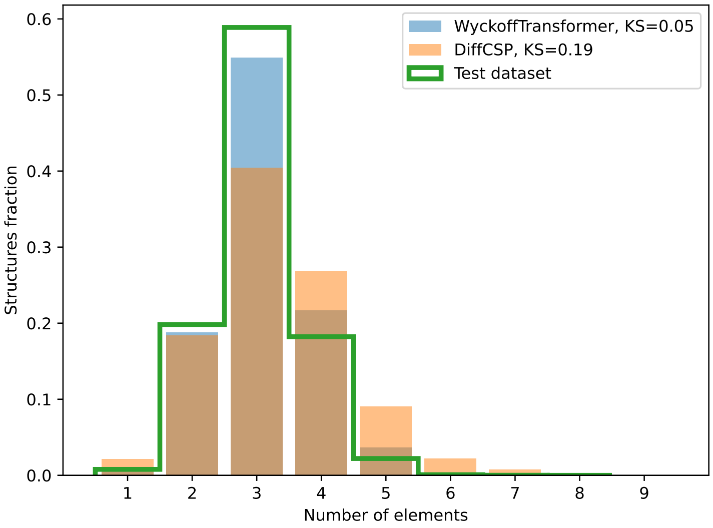

Figure 5: Distribution of the number of unique chemical elements per structure.

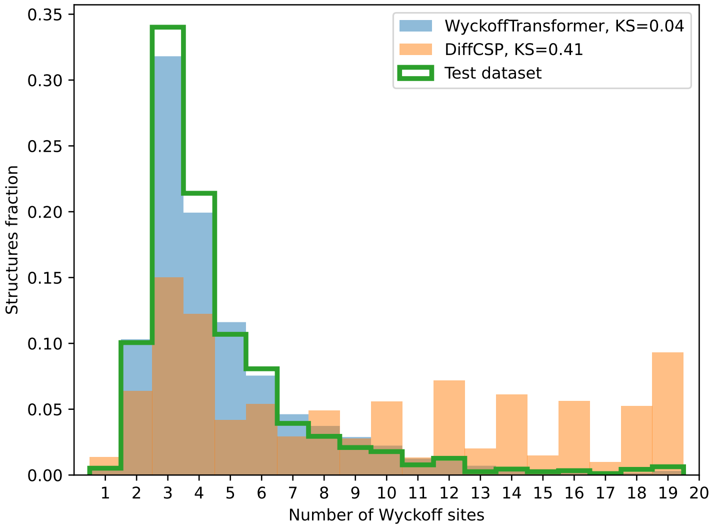

Figure 6: Distribution of the number of occupied Wyckoff positions per structure.

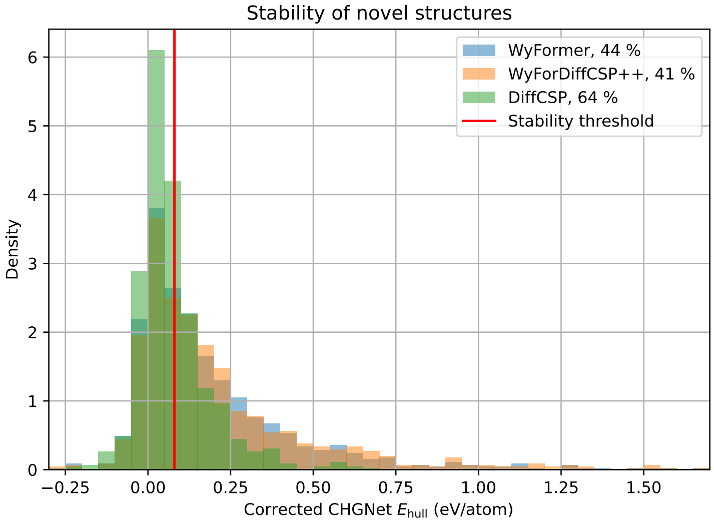

Figure 7: Stability of novel structures generated by WyckoffTransformer and DiffCSP trained on MP–20 dataset. Note that the % is not S.U.N., but just the fraction of novel structures that are stable.

C.2 WyckoffTransformer vs WyCryst; MP-20 binary and ternary

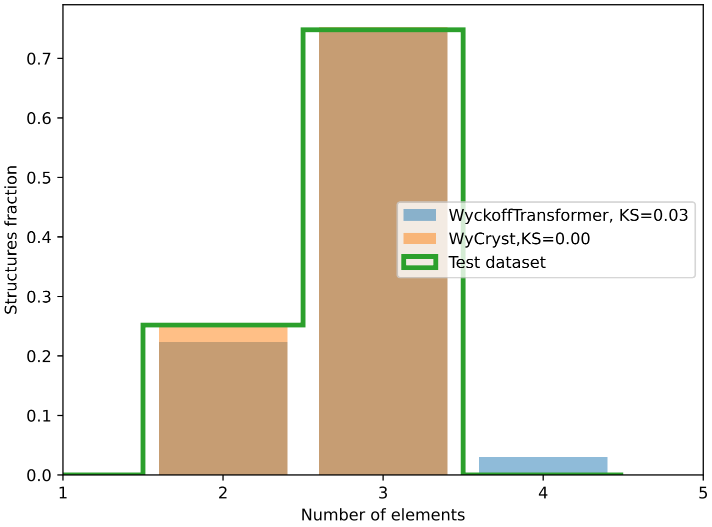

Figure 8: Distribution of the number of unique chemical elements per structure.

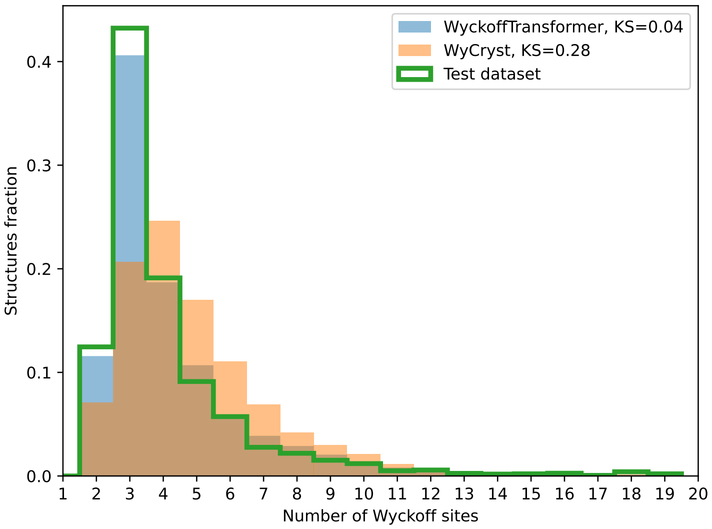

Figure 9: Distribution of the number of occupied Wyckoff positions per structure.

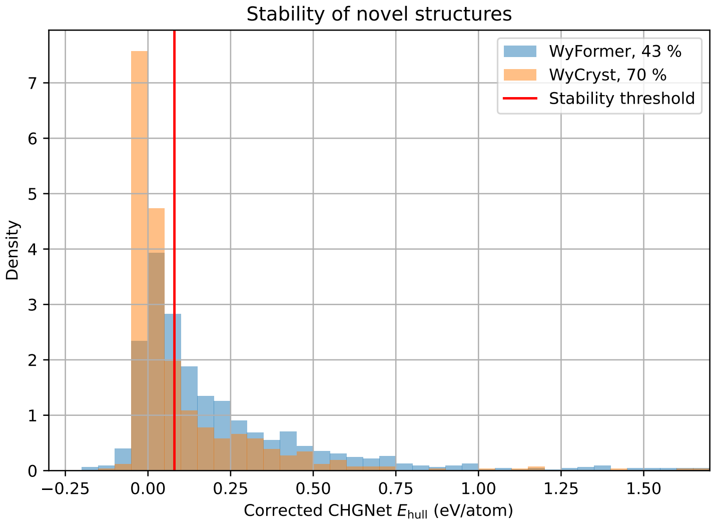

Figure 10: Stability of novel structures generated by WyCryst and WyckoffTransformer trained on a subset of MP-20 containing only binary and ternary structures. Note that the % is not S.U.N., but just the fraction of novel structures that are stable. Only half of the structures generated by WyCryst are novel, but the rest are rather stable – probably because they are similar to the structures in the training dataset.

D Permutation invariance

To asses learned permutation invariance, we compute the Kullback–Leibler divergence:

$$\mathrm{KL}\left[p_\theta(x)||p_\theta\left(\rho(x)\right)\right], \tag{1}$$

where x is a token sequence, $\rho(x)$ is a random permutation of the token sequence, $p_\theta(x)$ is the likelihood of x as computed by the model.

$$p_\theta(x) = p(t_0|\mathrm{SG})p(t_1|t_0, \mathrm{SG})...p(t_n|t_0, ..., t_{n-1}, \mathrm{SG}), \tag{2}$$

where t_i is the i-th part of token and SG is the space group, and the conditional probabilities are estimated by Wyckoff Transformer. Later we omit θ for clarity. To estimate $\mathbb{E}_{x\sim p(x)}$ we sample 9046 generated sequences; for each sequence, we sample 20 random permutations ($\mathbb{E}_\rho$):

$$\begin{aligned}\mathrm{KL} &= \mathbb{E}_{x\sim p(x)}\left[\log p(x) - \log p(\rho(x))\right] \\ &= \mathbb{E}_{x\sim p(x)}\left[\log p(x) - \mathbb{E}_\rho \log p(\rho(x))\right] \\ &= \mathbf{2.0}\end{aligned} \tag{3}$$

For a perfectly invariant model $\mathrm{KL} = 0$. For comparison, Shannon's entropy [18]:

$$H = -\mathbb{E}_{x\sim p(x)} \log p(x) = \mathbf{19.0}. \tag{4}$$

Finally, to get more comparison values, we compute the mean and standard deviation of the log-likelihoods over several datasets, see table 2. All generated samples are conditioned on the same space groups as the test dataset.

In conclusion, the model is not perfectly permutation invariant, but it is to a high degree. In terms of information theory, permutation constitutes 2 out of 19 *nats* needed to describe the state of the random variable (crystal in our case). In terms of naive log–likelihood arithmetic, standard deviation of log–likelihood between different permutations is two orders of magnitude less that the difference between log–likelihood of the generated and training data and 4 times less than the standard deviation of the generated samples log–likelihood.

Dataset	$-\mathbb{E}_{x\sim\text{dataset}}$	$\sigma_{x\sim\text{dataset}}$	$\mathbb{E}_{x\sim\text{dataset}}$
	$\log p(x)$	$\log p(x)$	$\sigma_\rho \log p(\rho(x))$
Generated, temperature = 100	499	373.6	16.8
Test	93.0	49.2	5.4
Train	92.8	48.3	5.5
Generated, temperature = 1	17.0	6.2	1.5
Generated, temperature = 10^{-2}	12.9	6.9	1.5

Table 2: Log-likelihood statistics of different datasets

E Structure generation details

The process begins by specifying a space group and defining WPs. PyXtal [7] allows users to input atomic species, stoichiometry, and symmetry preferences. Based on these parameters, PyXtal will generate a crystal structure that respects the symmetry requirements of the space group. The software is particularly useful for generating random structures that adhere to crystallographic symmetry rules: once the initial structure is generated, we then perform energy relaxation using CHGNet. CHGNet [6] is a neural network-based model designed to predict atomic forces and energies, significantly speeding up calculations that would traditionally require density functional theory (DFT). Energy relaxation involves optimizing the atomic positions to reach a minimum energy configuration, which represents the most stable form of the material. CHGNet, trained on vast DFT datasets, can efficiently relax crystal structures by adjusting atomic positions to reduce the total energy. This approach ensures that the final structure is not only symmetrical but also physically realistic in terms of energy stability.