WyckoffTransformer: Autoregressive Generation of Crystals

Nikita Kazeev[1], Andrey Ustyuzhanin[1,2], Ignat Romanov[3], Ruiming Zhu[4,5], Wei Nong[4], Shuya Yamazaki[4,5], Kedar Hippalgaonkar[4,5]
[1]Institute for Functional Intelligent Materials, National University of Singapore, Block S9, Level 9, 4 Science Drive 2, Singapore 117544
[2]Constructor University Bremen gGmbH, Campus Ring 1, Bremen, 28759, Germany
[3]HSE University, Myasnitskaya Ulitsa, 20, 101000, Moscow, Russia
[4]School of Materials Science and Engineering, Nanyang Technological University, Singapore 639798
[5]Institute of Materials Research and Engineering, Agency for Science, Technology and Research (A*STAR), Singapore 138634

kazeevn@gmail.com

Designing new materials is the ultimate goal of material science, an applied craft. Traditionally it's accomplished by proposing and then screening candidates. Computing the basic properties of a single crystal using ab initio methods, such as density functional theory (DFT), in the best case, requires several hours of computational time [1]. And even if a faster screening method is used, the space of all possible materials is intractable [2]. A better approach is to take advantage of the fact that stable crystals occupy only a small subspace of all possible atom combinations – and explore this space with generative models. We propose WyckoffTransformer, a machine learning model for generating crystals.

Our work relies on a crucial insight: most of the experimentally observed crystalline materials have internal symmetry, beyond unit cell transition, as shown in Figure 1. Those symmetries define optical, electrical, and magnetic properties [3]. For example, piezoelectric effects only appear in crystal classes that lack a center of symmetry [4]. Knowing just the symmetrical properties and chemical composition allows predicting the potential energy with accuracy comparable to a prediction based on complete structural information [5].

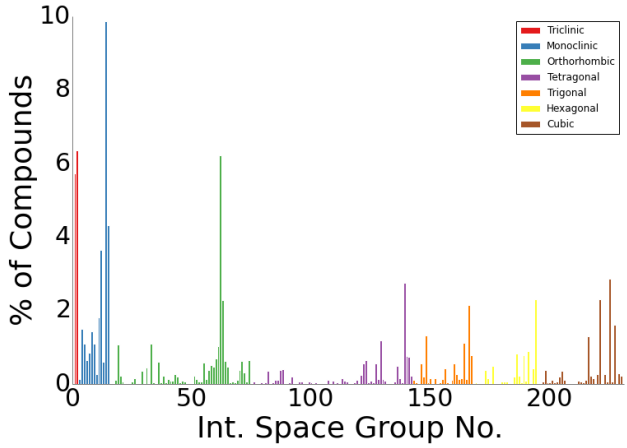

Figure 1: Distribution of symmetry groups in the Materials Project database [9,11]. Space group number greater than 1 indicates presence of symmetry beyond lattice translation.

A crystal is described by a symmetry group, which contains all transformations under which it is invariant. Space is separated into so-called Wyckoff positions, subspaces invariant under different symmetry operations. Each Wyckoff position can have 0 – 3 degrees of freedom. If an atom occupies a Wyckoff position, it's repeated 1 – 192 times across the unit cell, depending on the space group and position.

A crystal can be represented as a space group and a list of Wyckoff positions and elements occupying them. Such representation does not always completely define the structure, but greatly reduces the number of degrees of freedom. For the example in Figure 2, the reduction is from $26[\text{atoms}]\times 3[\text{coordinates}]+6[\text{lattice}]=84$ to just 4 (Wyckoff positions i and j each have a free parameter, and the lattice has two).

```
Composition: Nd2Al16Cu8
Group: I 4/m m m (139)
  8.9013,   8.9013,   5.1991,  90.0000,  90.0000,
90.0000, tetragonal
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]
```

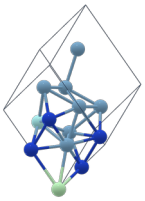

Figure 2: Wyckoff representation of $Nd(Al_2Cu)_4$ (mp-974729)

Such discrete representation naturally calls for a token-based model. The elements and Wyckoff sites can be ordered by electronegativity and then Wyckoff letter, thus allowing for an autoregressive approach. If we denote the space group as S, i-th element as e_i, site symmetry as s_i, and Wyckoff letter as w_i, then the distribution of valid structures can be factorized using the chain rule as a product of conditional probabilities:

$$P(S,e_{1..n},s_{1..n},w_{1..n})=p(S)\times p(e_1|S)\times$$
$$\times p(s_1|e_1,S)\times p(l_1|s_1,e_1,S)\times\ldots\times$$
$$\times p(l_n|s_n,e_n,l_{n-1},s_{n-1},e_{n-1},\ldots,S)$$

We use a Transformer Encoder [6] model to learn the conditional probabilities. Each token is a tuple (e_i,s_i,w_i,S) with each component being independently embedded and those embeddings concatenated. To learn the inter-token dependencies, we mask the corresponding parts of the token.

Related work Our work is a natural continuation of [5], the first generative model to utilize Wyckoff positions. The primary development 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 [7] independently of us explores a similar approach. The main difference is that we rely on site symmetries as opposed to

Wyckoff letters. Site symmetries are symmetry operations defined independently of the symmetry groups, thus allowing for a greater generalizability potential. We also use encoder with masking as opposed to decoder.

Results To evaluate WyckoffTransformer we trained it on the 27136 structures in the training part of the MP-20 dataset [8,9], using the validation part, 9047 structures, for early stopping and calibrating the sampling temperature. Then we used space groups and the first tokens of structures in the test part of the dataset to generate 9046 structures. 82% of the Wyckoff representations our model produced are valid and can be used to produce valid crystal structures under the validity metric commonly used to evaluate material generative models [8].

To evaluate how well the model learns the data distribution, we compare the distribution of high-level statistics between the generated and test data, depicted in Figures 3 and 4.

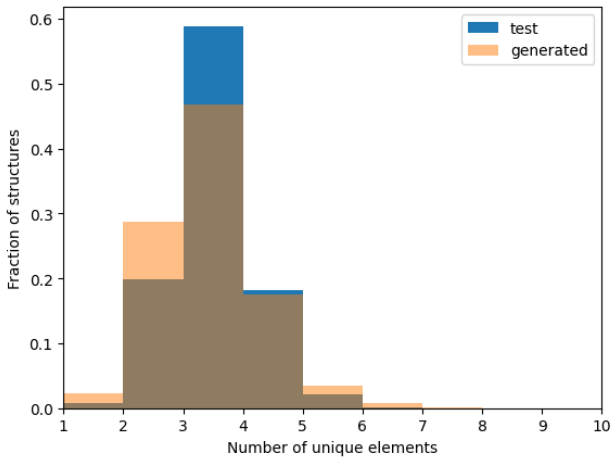

Figure 3: Distribution of the number of unique elements in the generated data and test part of the MP-20 dataset

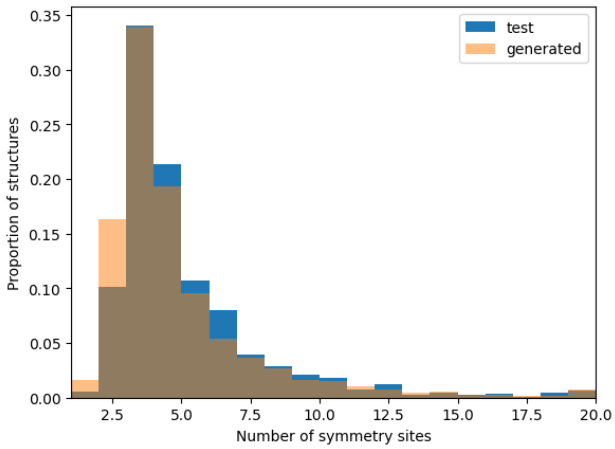

Figure 4: Distribution of the number of Wychoff sites in generated data and test part of the MP-20 dataset

We are working on evaluating the actual stability of produced structures using DFT.

Conclusion. We propose an advanced flexible token-based machine learning model for generating novel materials that takes advantage of the symmetry of nature. During evaluation, 82% of the produced structures were valid, with good correspondence of the high-level statistics between the generated and test data. Our work naturally complements diffusion-based approaches [10], by defining the high-level crystal structure and reducing the number of degrees of freedom for a later relaxation.

References

[1] Mardirossian, Narbe, and Martin Head-Gordon. Molecular physics 115.19 (2017): 2315-2372.
[2] D. W. Davies, K. T. Butler, A. J. Jackson, A. Morris, J. M. Frost, J. M. Skelton, and A. Walsh, Chem 1, 617 (2016).
[3] Malgrange, Cécile, Christian Ricolleau, and Michel Schlenker. Symmetry and physical properties of crystals. Springer, 2014.
[4] Yang, Jiashi. An Introduction to the Theory of Piezoelectricity. Vol. 9. Manhattan, New York: Springer, 2005
[5] Ruiming Zhu, Wei Nong, Shuya Yamazaki, Kedar Hippalgaonkar, arXiv:2311.17916 [cond-mat.mtrl-sci]
[6] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Advances in neural information processing systems 30 (2017).
[7] Cao, Z., Luo, X., Lv, J., & Wang, L. (2024). arXiv:2403.15734 [cond-mat.mtrl-sci]
[8] Xie, Tian, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. ICRL 2022 arXiv:2110.06197 [cs.LG]
[9] Jain, Anubhav, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia et al. APL materials 1, no. 1 (2013).
[10] Jiao, Rui, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. ICLR 2024 arXiv:2402.03992 [cs.LG]
[11] https://github.com/materialsvirtuallab/nano106