Crystal Structure Generation Based On Material Properties

Chao Huang^{1,2,a,b}, JiaHui Chen^{1,b}, HongRui Liang^{b,c}, ChunYan Chen^b, and Chen Chen^b

^aInstitute of Computing Technology, Chinese Academy of Science, Beijing, China ^bNingbo Institute of Information Technology Application, Chinese Academy of Sciences (CAS), Ningbo, China ^cZhejiang University, Hangzhou, China

The discovery of new materials is very important to the field of materials science. When researchers explore new materials, they often have expected performance requirements for their crystal structure. In recent years, data-driven methods have made great progress in the direction plane of crystal structure generation, but there is still a lack of methods that can effectively map material properties to crystal structure. In this paper, we propose a Crystal DiT model to generate the crystal structure from the expected material properties by embedding the material properties and combining the symmetry information predicted by the large language model. Experimental verification shows that our proposed method has good performance.

1 Introduction

Material science plays a crucial role in the development of modern technology and industrial production, with high-performance materials serving as the foundation for the manufacture of various advanced equipment. The generation of crystal structures is a central process driving the advancement of material scienceYao et al. (2023). As periodic materials, crystals are widely used in many important fields, including catalysts, alloys, and molds.

In recent years, data-driven methods have made great progress in the task of crystal structure generation (Nouira et al. (2018); Hoffmann et al. (2019); Hu et al. (2020); Ren et al. (2022)). Among various methods, diffusion model-based methods have been

¹ These authors have made equal contributions

² Corresponding author: chuang@ict.ac.cn

shown to be particularly effective in generating realistic and diverse crystal structures (Xie et al. (2021); Jiao et al. (2024a); Jiao et al. (2024b); Ye et al. (2024)). These methods use random processes to gradually transform random initial states into stable distributions, effectively capturing the complex landscape of crystal structures. On the other hand, methods based on autoregressive models have also achieved good results in generating crystal structures (Taniai et al. (2024)). These methods treat crystal structure data as strings and perform structure prediction in an autoregressive manner.

Despite the success of existing methods, few methods can accurately achieve end-to-end mapping of crystal properties to crystal structure. In this paper, we establish an end-to-end mapping between crystal properties and crystal structure through constraints on material properties and space groups. The method we proposed is called Uni-MDM, a universal material structure design model. This method establishes the relationship between material properties, space groups and crystal structures by combining the Fine-tuned GLM4(GLM et al. (2024)) model and the Crystal structure DiT model. Our contributions can be summarized as follows:

- 1. We divide the entire crystal structure generation process into two parts: first, the space group information is generated according to the required material properties, which is completed by the GLM4 model; second, the crystal structure is generated based on the material properties and space group information, which is completed by the DiT model.
- 2. We fine-tuned the GLM4 model through prompt engineering, allowing the model to output a reasonable number of crystal space groups and wyckoff positions based on the input elemental composition and material properties.
- 3. We proposed a DiT model that includes symmetry information constraints. By introducing material property embedding and crystal graph structure Transformer, the model can generate expected crystal structures through the constraints of material properties and space groups.
- 4. All our models have been trained and adapted on the NVIDIA platform and the Ascend Altas 800T A2 platform. Experiments show that our proposed method can generate stable crystal structures that meet the expected performance requirements under the constraints of material properties and space groups.

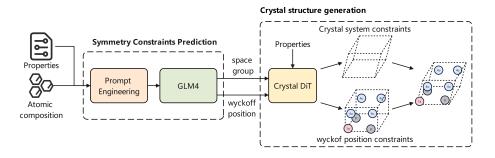


Figure 1: Uni-MDM.

2 Related works

Diffusion models. Diffusion modeling is a powerful generative model that generates data by simulating a gradual process of introducing noise and then learning how to reverse the process (Yang et al. (2024)). This approach has yielded significant results in several computer vision tasks such as image generation, image super-resolution, and image restoration. Subclasses of diffusion models include denoising diffusion probabilistic model (DDPM), noisy conditional score network (NCSN), and stochastic differential equation (SDE). The DDPM (Ho et al. (2020)) approach involves a forward propagation and a backward propagation. The forward process can be concluded as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t \mathbf{I}), \tag{1}$$

the backward process can be concluded as:

$$p_{\theta}(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_{\theta}(x_t, t), \Sigma_{\theta}(x_t, t)), \tag{2}$$

and the loss function is:

$$L(\theta) = \mathbb{E}_{x_0, \epsilon, t} \left[\frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1 - \bar{\alpha}_t)} ||\epsilon - \epsilon_{\theta}(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t)||^2 \right], \tag{3}$$

where \mathcal{N} denotes the normal distribution, β_t is the noise level parameter, α_t and $\bar{\alpha}t$ are the coefficients associated with the noise level, ϵ is the noise, ϵ_{θ} is the noise predicted by the model, and μ_t and Σ_t are the mean and covariance predicted by the inverse process, respectively.

The NCSN (Song and Ermon (2019)), is also utilized in the crystal structure generation task. The diffusion process and loss (Song and Ermon (2020)) can be summarized as follows:

$$p(x_t|x_0) = \mathcal{N}_w\left(x_t|x_0, \sigma^2 \mathbf{I}\right),\tag{4}$$

$$\frac{1}{2L} \sum_{i=1}^{L} \mathbb{E}_{p_{\text{data}}(x_0)} \mathbb{E}_{p_{\sigma_i}(x_t|x_0)} \left[\left\| \sigma_i \mathbf{s}_{\boldsymbol{\theta}}(x_t, \sigma_i) + \frac{x_t - x_0}{\sigma_i} \right\|_2^2 \right], \tag{5}$$

where $\mathbb{E}_{p_{\text{data}}(x_0)}$ denotes the expectation under the data distribution p_{data} , i.e., averaged over all possible data samples x_0 , and $\mathbb{E}_{p_{\sigma_i}(x_t|x_0)}$ denotes the expectation under the noise distribution p_{σ_i} given a data sample x_0 , i.e., averaging over all possible noise perturbations x_t . In addition, σ_i denotes the i-th noise level. As for the $\mathbf{s}_{\theta}(x_t,\sigma_i)$, it denotes the score network, parameterized by θ , which estimates the score (gradient) given the noisy data x_t and the noise level σ_i . The goal of this loss function is to train the score network \mathbf{s}_{θ} to be able to accurately estimate the gradient of the data distribution so that the original data can be recovered from the noise during the generation process. By minimizing this loss function, NCSN can learn how to generate high quality samples from noisy data.

Large language model. Traditional language model like the T5 (Raffel et al. (2020)) and BERT (Devlin et al. (2018)) models have demonstrated certain performance in the field of crystallography. Several vertical domain models based on T5 and BERT (Rubungo et al. (2023);Das et al. (2023)) have been used to encode crystal structure

data (e.g., atom types, coordinates, and space groups) into textual format and to predict crystal properties. However, these methodologies are currently limited to unidirectional predictions of properties with known crystal structures, lacking the capability to integrate crystallographic prior knowledge with the generative prowess of LLMs to inversely generate rational and diverse crystal structures. In recent years, with the rise of large language models and the application of pre-training-fine-tuning methods, the ability to understand general-purpose large language models has shown potential in the materials domain vertical. MatChat Chen et al. (2023) is an LLM dedicated to materials synthesis, fine-tuned on the basis of the LLaMA2-7B model. The model was trained on a dataset of over 13,000 high-confidence inorganic material synthesis pathways extracted from millions of scientific papers. MatChat demonstrates the ability to generate and reason about material synthesis knowledge, and shows excellent performance in predicting the synthesis of complex inorganic materials compared to ChatGPT. Nevertheless, there are fewer model designs based on Chinese generalized llm such as Glm, Qwen, Baichuan, etc., which are still under further research.

Data-driven crystal generation model. In the field of data-driven crystal generation models, researchers have developed a variety of generation models based on different data representations. These models include the use of atomic species with sites of the lattice (Sotskov et al. (2024)), voxel representations and distance matrices (Sultanov et al. (2023)), and 3D coordinates (Zhao et al. (2023)). Recently, the CDVAE (Sultanov et al. (2023)) model combines a VAE and a diffusion-based decoder to generate atoms and coordinates on a multigraph structure. The DiffCSP (Jiao et al. (2024a)) model optimizes the lattice matrices and atomic coordinates through a diffusion framework and applies polar decomposition to represent the symmetry of the lattice. Its improved version DiffCSP++ (Jiao et al. (2024a))introduces the symmetry information of the space group into the model and achieves better results. The PGCGM (Zhao et al. (2023)) model takes the space-group affine matrices as additional inputs but is limited by the ternary system. The PCVAE (Bao et al. (2022)) model, on the other hand, predicts the lattice parameters via conditional VAE and imposes lattice constraints in logarithmic space while specifying Wyckoff position constraints for all atoms. CrystalFormer model (Taniai et al. (2024)) is a data-driven crystal generation model that combines space group symmetry and autoregressive Transformer, which can rapidly generate possible crystal structures and provide diverse and stable initialized structures for existing crystal structure prediction software. Additionally, CrystaLLM (Antunes et al. (2024)) has been trained on a comprehensive dataset of millions of CIF files and is able to reliably generate correct CIF syntax and plausible crystal structures for many classes of inorganic compounds. These models demonstrate the potential of data-driven approaches to crystal structure generation and advance the field.

3 Our method

Crystal symmetry information is very important in the description of crystal geometry. a part of methods consider reference space group and wyckoff position as input when generating crystal structure, and get good results. However, for some crystal structures with high symmetry, the fixation of symmetry information can already determine most of the crystal parameters. Moreover, compared with establishing the connection between symmetry information and crystal structure, establishing the

connection between material properties and crystal structure is more in line with the research habits of researchers when exploring new crystal structures. Therefore, this paper attempts to generate crystal structure from material properties by combining a fine-tuned large language model for symmetric information prediction and a crystal structure DiT model.

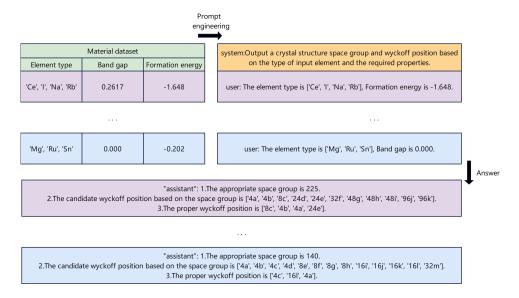


Figure 2: Prompt engineering to generate crystal symmetry information.

3.1 Large language model for crystal symmetry information prediction

The open source GLM4 large language model is already close to the world's strongest models Gemini Ultra and GPT-4. Its GLM4-9B version has shown superior performance over Llama-3-8B in semantic, mathematical, reasoning, code, knowledge and other aspects of the data set evaluation, and has better multilingual learning ability. Therefore, we chose GLM4-9b as the base model for training.

Since the GLM4-9B model has been pre-trained on a large number of text data, including the text of materials science, it has already understood a lot of basic knowledge of materials science, including the characteristics of individual atoms, the characteristics of spatial groups in crystal structures, and the meaning of wyckoff positions.

Therefore, we considered using the prompt engineering to create datasets and fine-tune the GLM4-9B model to allow the large language model to learn the correspondence between atomic properties, material properties and crystal symmetry information. The model prompt word is "output a crystal structure space group and wyckoff position based on the type of input element and the required properties." Then, the input text information contains atomic properties and required material properties, and the output answer generates the required result in a way similar to the chain of thought: first find the most suitable space group, then get the candidate wyckoff position based on the space group, and finally select the appropriate wyckoff position.

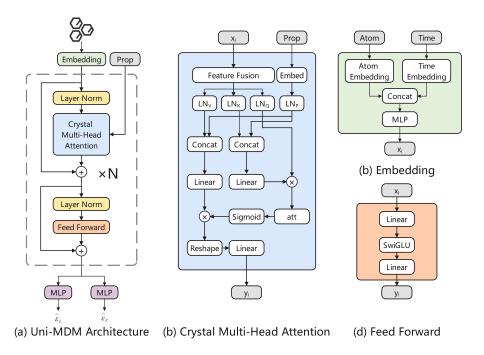


Figure 3: Overview of DiT structure.

3.2 Crystal structure diffusion model with transformers

In this section, we mainly introduce the designed crystal structure DiT model, which is responsible for receiving the symmetry information output from the large language model and generating the expected crystal structure according to the initial input atoms and material properties. The overall structure of the model is shown in Figure 3.Our model is mainly modified from the DiffCSP++Jiao et al. (2024b) model. This method mainly introduces symmetry information in the decoding part of the model to make the crystal structure the constraints of space group and wyckoff position after diffusion. We followed this idea and the edge feature extraction method of this method, and improved the overall network's ability to fit lattices and element coordinates by adding a crystal transformer block to the decoding module of the neural network. In particular, we designed a crystal multi-head attention module to fuse crystal features and expected material properties. The feature fusion module designed as

$$F_{ij}^{e} = \varphi_m(h_i, h_j, \varphi_h(k, \psi_{FT}(f_j - f_i)))$$
(6)

$$F^{c} = F^{n} + \varphi_{c}(F^{n}, \sum_{j=1}^{N} F_{ij}^{e})$$
(7)

where φ_m and φ_h are MLPs,and $\psi_{FT}: (-1,1)^3 \to [-1,1]^{3\times K}$ is the Fourier transformation with K bases relative fractional coordinate $f_j - f_i$. K is the unique O(3)-invariant representation of L. F_{ij}^e , F^n and F^c Respectively represent the extracted edge features, node features and crystal features of the network. The node features are obtained by the embedding module in Figure 3(b). h_i and h_j calculated from node features. The attention module design is as

$$head_i = Attention(F^c W_i^Q, \varphi_k(F^c W_i^K, P_{emb} W_i^P), \varphi_v(F^c W_i^V, P_{emb} W_i^P)))$$
(8)

Table 1: Crystal symmetry information prediction.

MP(2024.10.12))			
	Rouge-1	Rouge-2	Rouge-l
GLM4-9B(BG)	84.91	78.78	86.30
GLM4-9B(FM)	85.30	79.00	75.28

$$MultiHead(F^c, P_{emb}) = (head_1, ..., head_h)W^O$$
(9)

where φ_k and φ_v are linear layer, and P_{emb} represents the material properties after MLP embedding.

4 Experiments

In this paper, we trained and tested two material properties: band gap and formation energy. In the large language model fine-tuning, we fine-tuned both performances in a separate way, and the fine-tuned models are called GLM4-9B(BG) and GLM4-9B(FM). In Crystal DiT, we trained two versions of the model, the Crystal DiT version with the embedding band gap as a condition is Crystal DiT(BG), and the version with the embedding formation ability as a condition is Crystal DiT(FM).

Crystal symmetry information prediction. In order to fine-tune the GLM4-9B large model, we crawled 152,823 material datasets from the Materials Project (2024.10.12), and obtained the space group and wyckoff position of each material in the dataset through the Spacegroup Analyzer class in pymatgen and the pyxtal, and made a prompt engineering and constructed a new dataset in the way of subsection 3.1. After removing some data with too large atom numbers, the new dataset was constructed with a training test and validation set of 8:1:1, and the accuracy of the model-generated data was measured by Rouge-1, Rouge-2, and Rouge-1. To our knowledge, there are relatively few methods for generating symmetric information from material properties via autoregressive methods or large language models, so there is no baseline for comparison in this part.

Crystal structure generation. We use the Spacegroup Analyzer class of the pymatgen Ong et al. (2013) and the pyxtal to obtain the space group and wyckoff position of each material in the test dataset, and obtain the corresponding material properties. Inputting the above data into the Crystal DiT model, and get a crystal structure.

For each generated crystal structure, we calculate the match rate through the Structure-Matcher class in pymatgen, with thresholds stol=0.5, angle tol=10, ltol=0.3. The match rate represents the ratio of matched structures relative to the total number within the testing set, and the RMSD is averaged over the matched pairs, and normalized by $\sqrt[3]{V/N}$, where V is the volume of the lattice, N is the number of atom. We compared our approach with two classes of methods. The first class is the optimization-based methods including Random Search (RS), Bayesian Optimization (BO), and Particle Swarm Optimization (PSO). The second class considers three types of generative methods. P-cGSchNet (Gebauer et al. (2022)), CDVAE (Xie et al. (2021)) and DiffCSP++ (Jiao et al. (2024b)). The data for other methods are from DiffCSP++.

Table 2: Results on crystal structure prediction task. MR stands for Match Rate.

	MP-20	
	MR(%)	RMSE
RS	8.73	0.2501
ВО	8.11	0.2402
PSO	4.05	0.1567
P-cG-SchNet	15.39	0.3762
CDVAE	33.90	0.1045
DiffCSP++	80.27	0.0295
Crystal DiT(BG)	80.86	0.0376
Crystal DiT(FM)	81.48	0.0353

5 Conclusion

In this work, we propose Uni-MDM, a crystal generation method based on large language models and DiT, which effectively integrates material properties and space group constraints. We divide the entire crystal generation process into the prediction of symmetry information and the diffusion of crystal structure, and introduce material property constraints in both processes to ensure that the generated crystal structure meets expectations. Experimental results on the MP dataset show that our method has stable crystal structure generation results.

References

Luis M. Antunes, Keith T. Butler, and Ricardo Grau-Crespo. Crystal Structure Generation with Autoregressive Large Language Modeling, February 2024. URL http://arxiv.org/abs/2307.04340. arXiv:2307.04340 [cond-mat].

Xiaoyi Bao, Wang Zhongqing, Xiaotong Jiang, Rong Xiao, and Shoushan Li. Aspect-based Sentiment Analysis with Opinion Tree Generation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 4044–4050, Vienna, Austria, July 2022. International Joint Conferences on Artificial Intelligence Organization. ISBN 9781956792003. doi: 10.24963/ijcai.2022/561. URL https://www.ijcai.org/proceedings/2022/561.

Zi-Yi Chen, Fan-Kai Xie, Meng Wan, Yang Yuan, Miao Liu, Zong-Guo Wang, Sheng Meng, and Yan-Gang Wang. Matchat: A large language model and application service platform for materials science. *Chinese Physics B*, 32(11):118104, nov 2023. doi: 10.1088/1674-1056/ad04cb. URL https://dx.doi.org/10.1088/1674-1056/ad04cb.

Kishalay Das, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. Crysmmnet: Multimodal representation for crystal property prediction. *arXiv* preprint arXiv:2307.05390, 216:507–517, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pretraining of deep bidirectional transformers for language understanding. *arXiv* preprint arXiv:1810.04805, 2018.

- Niklas WA Gebauer, Michael Gastegger, Stefaan SP Hessmann, Klaus-Robert Müller, and Kristof T Schütt. Inverse design of 3d molecular structures with conditional generative neural networks. *Nature communications*, 13(1):973, 2022.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.
- 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. *arXiv preprint arXiv:1909.00949*, 2019.
- Jianjun Hu, Wenhui Yang, and Edirisuriya M Dilanga Siriwardane. Distance matrix-based crystal structure prediction using evolutionary algorithms. *The Journal of Physical Chemistry A*, 124(51):10909–10919, 2020.
- 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, 2024a.
- Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation. *arXiv preprint arXiv*:2402.03992, 2024b.
- Asma Nouira, Nataliya Sokolovska, and Jean-Claude Crivello. Crystalgan: learning to discover crystallographic structures with generative adversarial networks. *arXiv* preprint arXiv:1810.11203, 2018.
- 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.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140): 1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.
- Zekun Ren, Siyu Isaac Parker Tian, Juhwan Noh, Felipe Oviedo, Guangzong Xing, Jiali Li, Qiaohao Liang, Ruiming Zhu, Armin G Aberle, Shijing Sun, et al. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter*, 5(1):314–335, 2022.
- Andre Niyongabo Rubungo, Craig Arnold, Barry P Rand, and Adji Bousso Dieng. Llm-prop: Predicting physical and electronic properties of crystalline solids from their text descriptions. *arXiv* preprint arXiv:2310.14029, 2023.
- Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf.

- Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12438–12448. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92c3b916311a5517d9290576e3ea37ad-Paper.pdf.
- Vadim Sotskov, Evgeny V. Podryabinkin, and Alexander V. Shapeev. A machine-learning potential-based generative algorithm for on-lattice crystal structure prediction. *MRS bulletin*, (3):49, 2024.
- Arsen Sultanov, Jean-Claude Crivello, Tabea Rebafka, and Nataliya Sokolovska. Data-Driven Score-Based Models for Generating Stable Structures with Adaptive Crystal Cells. *Journal of Chemical Information and Modeling*, 63(22):6986–6997, November 2023. ISSN 1549-9596, 1549-960X. doi: 10.1021/acs.jcim.3c00969. URL https://pubs.acs.org/doi/10.1021/acs.jcim.3c00969.
- Tatsunori Taniai, Ryo Igarashi, Yuta Suzuki, Naoya Chiba, Kotaro Saito, Yoshitaka Ushiku, and Kanta Ono. Crystalformer: infinitely connected attention for periodic structure encoding. *arXiv preprint arXiv:*2403.11686, 2024.
- Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *arXiv* preprint arXiv:2110.06197, 2021.
- Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion Models: A Comprehensive Survey of Methods and Applications. *ACM Computing Surveys*, 56(4):1–39, April 2024. ISSN 0360-0300, 1557-7341. doi: 10.1145/3626235. URL https://dl.acm.org/doi/10.1145/3626235.
- Zhenpeng Yao, Yanwei Lum, Andrew Johnston, Luis Martin Mejia-Mendoza, Xin Zhou, Yonggang Wen, Alán Aspuru-Guzik, Edward H Sargent, and Zhi Wei Seh. Machine learning for a sustainable energy future. *Nature Reviews Materials*, 8(3):202–215, 2023.
- Cai-Yuan Ye, Hong-Ming Weng, and Quan-Sheng Wu. Con-cdvae: A method for the conditional generation of crystal structures. *Computational Materials Today*, 1:100003, 2024.
- Yong Zhao, Edirisuriya M. Dilanga Siriwardane, Zhenyao Wu, Nihang Fu, Mohammed Al-Fahdi, Ming Hu, and Jianjun Hu. Author Correction: Physics guided deep learning for generative design of crystal materials with symmetry constraints. *npj Computational Materials*, 9(1):104, June 2023. ISSN 2057-3960. doi: 10.1038/s41524-023-01059-8. URL https://www.nature.com/articles/s41524-023-01059-8.