A PROPERTY-PROMPTED MULTI-SCALE DATA AUGMENTATION APPROACH FOR CRYSTAL REPRESENTATION

Zhongyi Deng
South China University of Technology & Nanyang Technological University

Shuzhou Li*
School of Materials Science and Engineering
Nanyang Technological University,
Singapore 639798, Singapore

Tong Zhang* & C. L. Philip Chen
School of Computer Science and Engineering
South China University of Technology
Guangzhou 510006, China

ABSTRACT

The inverse design of crystals with multiple objectives represents a significant challenge in materials science. The interplay among various desired properties often results in unbalanced crystal structure generation. In schemes based on generative language models, this issue primarily stems from the models' limited capability to learn continuous property values, compounded by the scarcity of high-quality material data for training. To address these challenges, a property prompt-based scheme has been proposed to achieve multi-scale data augmentation for crystal representation. This scheme constructs learnable prompt templates for the single property and extends them to multiple properties. The property prompt introduces learnable templates that map continuous property values to discrete prompt spaces, enhancing the learning ability of generative language models for discrete property values. Multi-scale data augmentation disentangles the interactions between various material properties and transforms them into mutual promotion through end-to-end pre-training, thereby alleviating the problem of insufficient high-quality material data. The scheme has been validated for key properties that affect the crystal structure composition, including the formation energy and the band gap, as well as their various combinations. Experimental results demonstrate that the proposed model achieves significant performance improvements across multiple target property combinations, showcasing its robust representation and generalization capabilities in the inverse design of crystals with multiple objectives.

1 INTRODUCTION

Crystalline materials are fundamental to a wide range of applications, from semiconductors to catalysts. The ability to predict and design these structures with precision is critical to optimizing their propertiesFang et al. (2022); Lee et al. (2024). Traditional approaches to material discovery, often reliant on trial-and-error experimentation or computationally expensive quantum mechanical simulations, are increasingly being augmented by artificial intelligence (AI) techniquesZeni et al. (2025). These models, trained on vast datasets of material properties and structures, can generate novel crystal structures, predict material properties, and even suggest synthesis pathwaysMerchant et al. (2023). The AI-based material structure prediction relies on graph neural networks, which has shortcoming in computationOck et al. (2023). These limitations make it difficult to capture details, such as the subtle differences in the geometry and length of the interatomic bonds in the crystal and the resulting electronic and optical properties.

The LLM-based approach provides a fresh perspective on the design material problemSuvarna et al. (2023); Zheng et al. (2024); Lu et al. (2024); Choi & Lee (2024). LLM-Prop has been proposed

*Corresponding author:lisz@ntu.edu.sg, tony@scut.edu.cn

to address this challenge by shifting the focus from graphics to the text domainRubungo et al. (2023). This model, based on the T5 architecture and trained on a textual benchmark containing over 140,000 crystal descriptions, demonstrates the ability to represent crystal structures from text descriptions. Text-based methods hold significant potential in materials science. An example of this is DARWIN, a specialized LLM based on the LLaMA architecture, meticulously designed for applications in materials scienceXie et al. (2023). Its training involved a variety of instructions to ensure the factual accuracy of outputs and explored the interconnections between different scientific tasks through multi-task training strategies. The LLM-based approach has great advantages in automating the extraction of material knowledge from scientific literatureJia et al. (2024); Liu et al. (2024a); Polak & Morgan (2024). By training on large corpora of research articles, these models can identify trends, correlations, and potential material candidates that might otherwise remain undiscovered. One of the most significant breakthroughs in this field is the application of LLMs to encode and decode material representations. For instance, models like GPT-4 and its variants have been adapted to process and generate material descriptors, enabling the prediction of crystal structures from chemical formulas or vice versaPatel & Wong (2023); Zheng et al. (2023). Recent generative models have demonstrated significant potential; however, their adoption has been hindered by inefficiencies, architectural constraints, and limited open-source availability. The representation of crystal structures using the Simplified Line-Input Crystal-Encoding System (SLICES) notation as a string of characters enables the use of state-of-the-art natural language processing models for crystal designXiao et al. (2023); Wang et al. (2024). Inspired by the GPT models and the SLICES notation, MatterGPT is proposed, demonstrating powerful on-demand crystal generation capabilities and excelling in producing structures with specific attributesChen et al. (2024).

Despite these advancements, challenges remain in fully leveraging LLMs for material discovery. Issues such as data scarcity and the need for validation frameworks must be addressed to ensure the reliability and scalability of these approachesKang et al. (2025). The study employs the SLICES as a grammatical foundation to design an extensible prompt learning that aims to enhance the crystal representation capabilities of generative language models. The approach simultaneously considers various crystal structure properties in language modeling, achieving multi-scale data augmentation during pre-training. Independent modeling of single and multiple properties represents the optimal performance of LLM-based schemes under specific conditions and forms the experimental baseline. Experiments demonstrate that extensible property prompt learning can approach the optimal performance of independent modeling through a single pre-training session. The proposed model's ability to generate material structures under multi-properties even surpasses the baseline.

2 RELATED WORK

2.1 SLICES

The Simplified Line-Input Crystal-Encoding System (SLICES) is the first invertible and invariant crystal representation tool, supporting encoding and decoding crystal structures, reconstructing them, and generating new materials with desired properties using generative LLMsXiao et al. (2023).

2.2 GENERATIVE LANGUAGE MODEL

Drawing inspiration from the success of GPT models in generating coherent text, the MatterGPT was proposed and trained on the next-token prediction task to generate solid-state materials with targeted propertiesChen et al. (2024). MatterGPT's capability was demonstrated to generate de novo crystal structures with targeted single properties, including both lattice-insensitive (formation energy) and lattice-sensitive (band gap) properties.

2.3 PROMPT LEARNING

Prompt learning is a machine learning technique that leverages prompts to guide the model's outputsLi & Liang (2021); Liu et al. (2023; 2024b). In the context of material discovery, prompt learning involves using specifically designed prompts to direct a model's focus on generating or analyzing material properties and structuresLee et al. (2024). By integrating domain-specific knowledge into

the prompts, the model can be fine-tuned to predict material characteristics, discover new materials, and optimize existing onesLiu et al. (2024a; 2025).

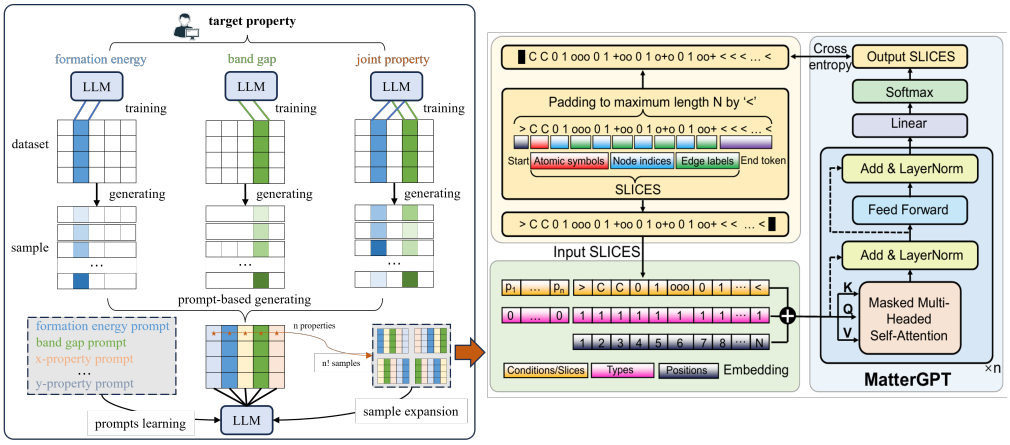
3 METHODOLOGY

3.1 PROPERTY PROMPTS

Prompt is the technique of making better use of the knowledge from the pre-trained model by adding additional texts to the input. As shown in Fig. 1 (a), each material property was classified with a specific prompt token. The language model simultaneously attend to multiple properties of the crystal structure samples through soft prompt learning in the training. In tasks aimed at generating different target properties, the prompt token for the target property activated specific knowledge within the language model, facilitating the generation of crystal samples with specific property. For instance, in this work, the language model focused on the formation energy and band gap of crystal structures during training. The target attribute values and their corresponding prompt tokens were input into the trained model, enabling the generation of crystal samples with formation energy, band gap, and their combination using the same set of model parameters.

3.2 SAMPLE EXPANSION

Given that language models can recognize different material properties through prompt tokens, the scale of the training dataset is expanded by rearranging the input order of these properties. This helps alleviate the limitation of insufficient high-quality data. Assuming n material properties are considered simultaneously, different permutations of these properties with the same crystal structure can form new samples. Ideally, the training sample size could be expanded by a factorial of n.



(a) data argumentation (b) model architecture

Figure 1: Data augmentation workflow based on properties prompts.

4 EXPERIMENTS

4.1 EXPERTMENT SETUP

4.1.1 DATA

The filtered Alex-20 dataset, utilized in this study, originates from the Alexandria databaseSchmidt et al. (2021; 2023). The filtering process in MatterGPT excludes crystals with atomic numbers exceeding 86Spicher & Grimme (2020); Xiao et al. (2023), structures with dimensions lower than 3Eon (2011), and all structures classified as metallic based on their electronic propertiesJung et al. (2024); Talapatra et al. (2023). The resulting filtered Alex-20 dataset comprises 280,033 unique crystal structures, ensuring high data quality and processability.

4.1.2 MODEL

The proposed approach was implemented in MatterGPT Chen et al. (2024) in which the necessary dependencies was provided by the SLICES package Xiao et al. (2023).

All baseline models are trained for 50 epochs using the Adam optimization algorithm, with an initial learning rate set at 0.0001. The training is conducted with a batch size of 60, an embedding size of 768, a total of 12 Transformer-decoder layers, and 12 attention heads per layer, resulting in approximately 80 million trainable hyperparameters.

All the hyperparameters, including learning rate, batch size, and number of layers, are meticulously tuned through a grid search to ensure optimal performance under different experimental conditions.

4.1.3 EXPERIMENTAL DESIGN

Experimental and Control Groups. Control Groups consist of independently trained models under different conditions of Independent Variables. Experimental Groups utilize pre-trained models based on the proposed scheme applied to the dataset. This means that Experimental Groups contain only one set of model parameters.

Independent Variables. In single-objective experiments, various formation energies and band gaps are used as independent variables, while in multi-objective experiments, sampled combinations of these attributes serve as independent variables.

Dependent Variable. This study builds upon the experimental setup of MatterGPT, using validity as the dependent variable for both experimental and control groups. Validity is defined as the proportion of valid samples among the generated samples. A generated string is considered a valid sample if it can be reconstructed into the original crystal structure. Higher validity indicates a model's improved capability to generate crystal structures conforming to SLICES syntax.

Control Variables. Control variables include a series of expected numbers of generated samples. Ten expected values ranging from 1000 to 10000 are set to observe changes in the model's generation capability with increasing sample size.

4.2 EXPERIMENT RESULTS

4.2.1 SAMPLE GENERATION UNDER SINGLE TARGET PROPERTY

To verify the model's ability to characterize single target property in crystals, different numbers of samples were generated for formation energy and bandgap. The statistical analysis of the filtered Alex dataset revealed that the formation energy of the training samples was concentrated in the [-1, -4] interval, and the bandgap of the training samples was concentrated in the [1, 4] interval.

Consequently, the target values for formation energy were set to [-1, -2, -3, -4] and for bandgap to [1, 2, 3, 4] to maximize the model's generation capability. Ten points were evenly selected from the [1000, 10000] interval as the target number of generated samples.

The experimental results for formation energy are shown in Fig. 2. The generation ability curve indicates that the model's performance under the formation energy condition is slightly weaker than that of the original model, with a performance decrease of less than 2%.

The experimental results for bandgap are shown in Fig. 3. The generation ability curve indicates that the model's performance under the bandgap condition is slightly better than that of the original model, with a performance increase of less than 1%.

These experimental results demonstrate that property-based prompt learning enables the same set of model parameters to approach or even surpass the performance of independently modeled parameters for different property.

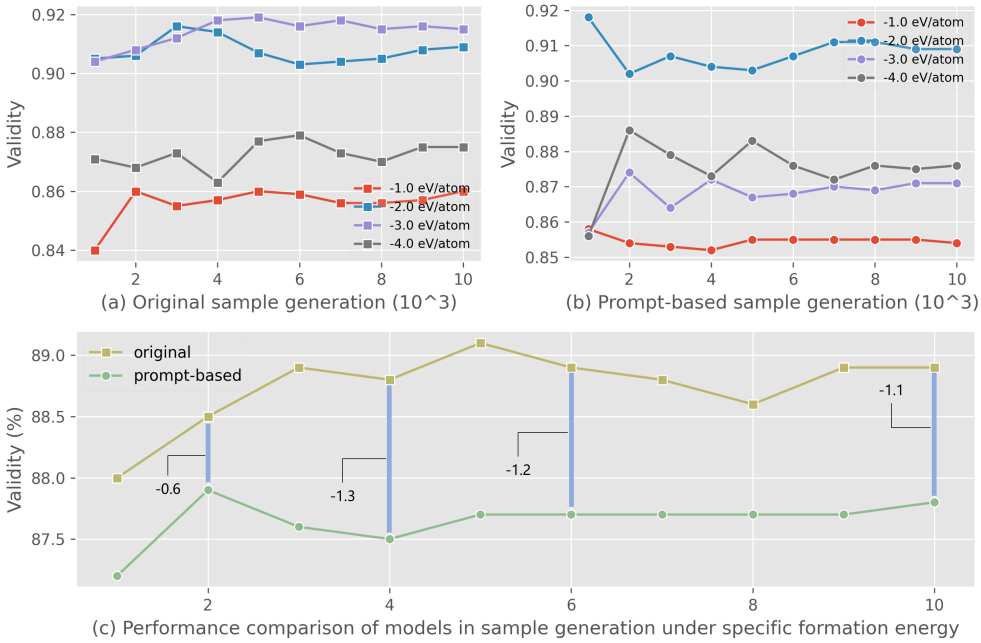


Figure 2: Sample generation under specific formation energy. Figures (a) and (b) display the validity of crystal structures produced by the original model and property-prompted model, respectively, under given formation energy conditions. Figure (c) compares the average validity between both models.

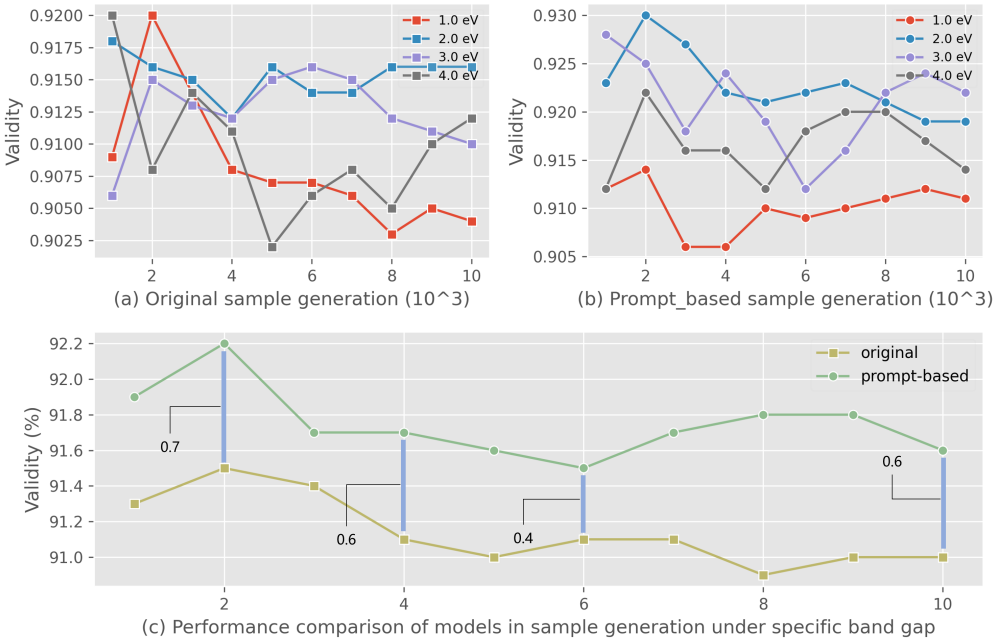


Figure 3: Sample generation under specific band gap. Figures (a) and (b) display the validity of crystal structures produced by the original model and property-prompted model, respectively, under given band gap conditions. Figure (c) compares the average validity between both models.

4.2.2 SAMPLE GENERATION UNDER MULTIPLE TARGET PROPERTIES

To validate the model's generation ability under multi-objective properties, different combinations of formation energy and bandgap were set. The target values for formation energy were set to [1, 0, -1, -2, -3, -4], and the target values for bandgap were set to [1, 2, 3, 4, 5, 6], to achieve a comprehensive evaluation of the model's crystal representation ability. According to the property distribution of the training dataset, the model's crystal representation ability is strongest for formation energies in the [-1, -4] interval and bandgaps in the [1, 4] interval. In other regions, due to insufficient training data, the model did not receive sufficient learning. The heatmap in Figure 4(a) of the original model substantiates this analysis. The effectiveness of sample generation in the formation energy [0, 1] region and the bandgap [5, 6] region is significantly weaker in the original model compared to regions with sufficient training samples. In contrast, the heatmap in Figure 4(b) of the property-prompted model demonstrates well-balanced and robust generation capabilities across all tested regions.

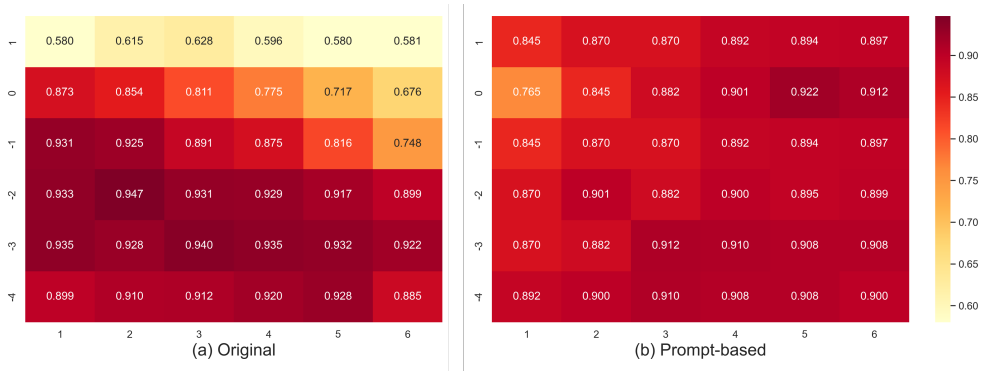


Figure 4: Sample generation under multiple target properties. The figure's horizontal axis shows the bandgap, while the vertical axis represents the formation energy. The grid values indicate the validity of the generated crystal structures under these dual conditions.

5 CONCLUSION

In this study, a property-prompted multi-scale data augmentation approach is proposed for the inverse design of material, which enhances the crystal representation capabilities of generative language models under specific properties. Multi-scale data augmentation is achieved during training of the generative language model based on different material properties. The proposed approach is implemented on the GPT architecture and validated under various property combinations, demonstrating the model's performance. The experimental results indicate that model parameters based on property prompts can approach or even surpass the performance of independently modeled parameters in single target property generation with just one training process, while also achieving more balanced performance in multi-objective property generation.

REFERENCES

Yan Chen, Xueru Wang, Xiaobin Deng, Yilun Liu, Xi Chen, Yunwei Zhang, Lei Wang, and Hang Xiao. Mattergpt: A generative transformer for multi-property inverse design of solid-state materials, 2024. URL https://arxiv.org/abs/2408.07608.

Jaewoong Choi and Byungju Lee. Accelerating materials language processing with large language models. *Communications Materials*, 5(1):13, 2024.

J-G Eon. Euclidian embeddings of periodic nets: definition of a topologically induced complete set of geometric descriptors for crystal structures. *Acta Crystallographica Section A: Foundations of Crystallography*, 67(1):68–86, 2011.

Jiheng Fang, Ming Xie, Xingqun He, Jiming Zhang, Jieqiong Hu, Yongtai Chen, Youcai Yang, and Qinglin Jin. Machine learning accelerates the materials discovery. *Materials Today Communications*, 33:104900, 2022.

Shuyi Jia, Chao Zhang, and Victor Fung. Llmatdesign: Autonomous materials discovery with large language models. *arXiv preprint arXiv:2406.13163*, 2024.

Son Gyo Jung, Guwon Jung, and Jacqueline M Cole. Automatic prediction of band gaps of inorganic materials using a gradient boosted and statistical feature selection workflow. *Journal of Chemical Information and Modeling*, 64(4):1187–1200, 2024.

Yeonghun Kang, Wonseok Lee, Taeun Bae, Seunghee Han, Huiwon Jang, and Jihan Kim. Harnessing large language models to collect and analyze metal–organic framework property data set. *Journal of the American Chemical Society*, 2025.

Namkyeong Lee, Heewoong Noh, Sungwon Kim, Dongmin Hyun, Gyoung S Na, and Chanyoung Park. Density of states prediction of crystalline materials via prompt-guided multi-modal transformer. *Advances in Neural Information Processing Systems*, 36, 2024.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021.

Hongxuan Liu, Haoyu Yin, Zhiyao Luo, and Xiaonan Wang. Integrating chemistry knowledge in large language models via prompt engineering. *Synthetic and Systems Biotechnology*, 10(1):23–38, 2025.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

Siyu Liu, Tongqi Wen, ASL Subrahmanyam Pattamatta, and David J Srolovitz. A prompt-engineered large language model, deep learning workflow for materials classification. *Materials Today*, 80:240–249, 2024a.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *AI Open*, 5:208–215, 2024b.

Muyu Lu, Fengyu Gao, Xiaolong Tang, and Linjiang Chen. Analysis and prediction in scr experiments using gpt-4 with an effective chain-of-thought prompting strategy. *Iscience*, 27(4), 2024.

Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

Janghoon Ock, Chakradhar Guntuboina, and Amir Barati Farimani. Catalyst energy prediction with catberta: Unveiling feature exploration strategies through large language models. *ACS Catalysis*, 13(24):16032–16044, 2023.

Dylan Patel and Gerald Wong. Gpt-4 architecture, infrastructure, training dataset, costs, vision, moe. *Demystifying GPT-4: The Engineering Tradeoffs That Led OpenAI to Their Architecture. SemiAnalysis*, 10:1–17, 2023.

Maciej P Polak and Dane Morgan. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15(1):1569, 2024.

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.

Jonathan Schmidt, Love Pettersson, Claudio Verdozzi, Silvana Botti, and Miguel AL Marques. Crystal graph attention networks for the prediction of stable materials. *Science advances*, 7(49): eabi7948, 2021.

Jonathan Schmidt, Noah Hoffmann, Hai-Chen Wang, Pedro Borlido, Pedro JMA Carriço, Tiago FT Cerqueira, Silvana Botti, and Miguel AL Marques. Machine-learning-assisted determination of the global zero-temperature phase diagram of materials. *Advanced Materials*, 35(22):2210788, 2023.

Sebastian Spicher and Stefan Grimme. Robust atomistic modeling of materials, organometallic, and biochemical systems. *Angewandte Chemie International Edition*, 59(36):15665–15673, 2020.

Manu Suvarna, Alain Claude Vaucher, Sharon Mitchell, Teodoro Laino, and Javier Pérez-Ramírez. Language models and protocol standardization guidelines for accelerating synthesis planning in heterogeneous catalysis. *Nature Communications*, 14(1):7964, 2023.

Anjana Talapatra, Blas Pedro Uberuaga, Christopher Richard Stanek, and Ghanshyam Pilania. Band gap predictions of double perovskite oxides using machine learning. *Communications Materials*, 4(1):46, 2023.

Baoning Wang, Zhiyuan Xu, Zhiyu Han, Qiwen Nie, Hang Xiao, and Gang Yan. Slices-plus: A crystal representation leveraging spatial symmetry, 2024. URL https://arxiv.org/abs/2410.22828.

Hang Xiao, Rong Li, Xiaoyang Shi, Yan Chen, Liangliang Zhu, Xi Chen, and Lei Wang. An invertible, invariant crystal representation for inverse design of solid-state materials using generative deep learning. *Nature Communications*, 14(1):7027, 2023.

Tong Xie, Yuwei Wan, Wei Huang, Zhenyu Yin, Yixuan Liu, Shaozhou Wang, Qingyuan Linghu, Chunyu Kit, Clara Grazian, Wenjie Zhang, et al. Darwin series: Domain specific large language models for natural science. *arXiv preprint arXiv:2308.13565*, 2023.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, et al. A generative model for inorganic materials design. *Nature*, pp. 1–3, 2025.

Zhiling Zheng, Oufan Zhang, Christian Borgs, Jennifer T Chayes, and Omar M Yaghi. Chatgpt chemistry assistant for text mining and the prediction of mof synthesis. *Journal of the American Chemical Society*, 145(32):18048–18062, 2023.

Zhiling Zheng, Federico Florit, Brooke Jin, Haoyang Wu, Shih-Cheng Li, Kakasaheb Y Nandiwale, Chase A Salazar, Jason G Mustakis, William H Green, and Klavs F Jensen. Integrating machine learning and large language models to advance exploration of electrochemical reactions. *Angewandte Chemie*, pp. e202418074, 2024.